



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

Prospettive di sviluppo dei sistemi di archiviazione documentale tra Cloud ed Intelligenza Artificiale

Matteo Cristani

Professore Associato di Sistemi di Elaborazione dell'Informazione



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**





L'ARCHIVIAZIONE DOCUMENTALE

- Un macrofenomeno, quasi una moda. L'archiviazione di qualsiasi documento;
- Dobbiamo forse imparare a distinguere ciò che va archiviato da ciò che va dimenticato?
- Un'idea di archivio è anche un'idea della *rilevanza* dei documenti.
- La rilevanza è un concetto *modulare*. Ciò che è rilevante per una categoria di utenti può essere irrilevante per un'altra.



PRIMA PARTE: Disegnare gli archivi

Perché progettare un archivio è un problema molto diverso da quello di progettare una base di dati? Un viaggio nell'arte del design documentale, tra biblioteche polverose e sistemi digitali, la «dittatura» dei bibliotecari, i vaneggiamenti misoneisti della carta migliore del digitale e le velleità del «non-progetto».



SECONDA PARTE: Archivio, ritrovo, riarchivio. Un'eterna ghirlanda brillante

Ritorniamo a ciò che era, e sempre dovrebbe essere, la missione di questa abbazia: la preservazione del sapere... preservazione ho detto...non ricerca del sapere... perché non c'è progresso nella storia della conoscenza, ma una mera, costante e sublime ricapitolazione.

Ven. Jorge da Burgos – Il nome della Rosa

Mai porremo la parola “fine” all’impegno intellettuale dell’uomo e la matematica non finirà. Mai ci sarà un ultimo risultato della matematica.

Emilio Ambrisi – Gödel, Escher, Bach: bellezza, eternità, insegnamento
(Periodico di Matematiche Gennaio 2019 n. 1-2, pag. 5-11.



TERZA PARTE: Chiedi alla polvere (o a ChatGPT?)

[...] e la biblioteca con i grossi nomi degli scaffali, il vecchio Dreiser, il vecchio Mencken, tutta la banda riunita che andavo a riverire. Salve Dreiser, ehi Mencken, ciao a tutti, c'è un posto anche per me nel settore della B, B come Bandini, stringetevi un po', fate posto ad Arturo Bandini. Mi sedevo al tavolo e guardavo verso il punto in cui avrebbero messo il mio libro, proprio lì, vicino ad Arnold Bennett; niente di speciale quell'Arnold Bennett, ma ci sarei stato io a tenere alto l'onore delle B, io, il vecchio Arturo Bandini, uno della banda.

John Fante – Chiedi alla polvere



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

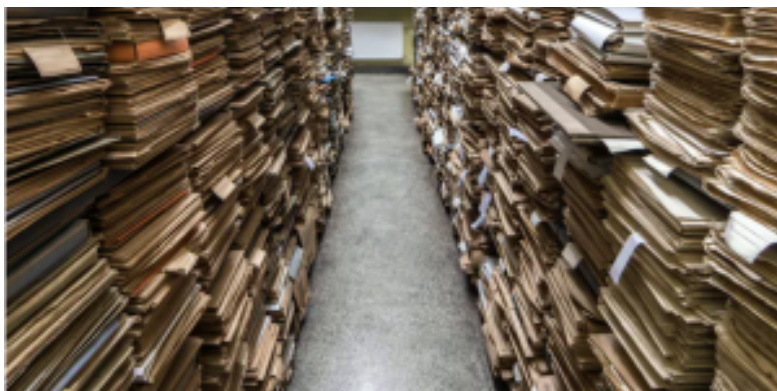
PRIMA PARTE: DISEGNARE GLI ARCHIVI



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

CATTIVO ARCHIVIO



BUON ARCHIVIO

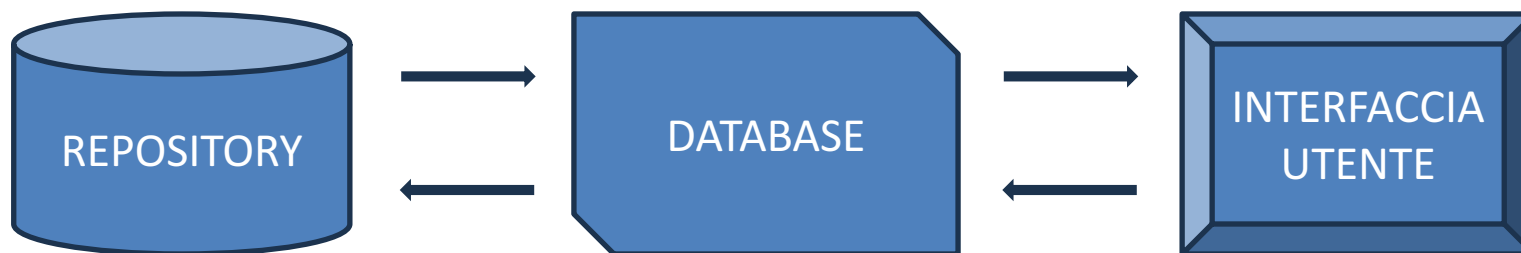




CHE COS'È UN ARCHIVIO

Un archivio è una struttura di gestione di dati formata da tre componenti chiave:

1. Il **repository** documentale;
2. La **base di dati** di gestione del repository;
3. Il sistema di **interfaccia utente**.





LE FUNZIONI GENERALI DI UN ARCHIVIO

1. Definire le modalità di accesso ai documenti, i metadati associati ai singoli documenti, il formato dei documenti e potenzialmente cambiare le stesse componenti (**design e redesign**);
2. Editare i metadati corrispondenti ad un singolo documento, producendo quindi l'archiviazione effettiva dei documenti (**metadata editing**);
3. Inserire un singolo documento, caricandolo sul sistema, ovvero posizionandolo nel repository (**inserimento in archivio**).



USE CASES

- Un utente con poteri di **design e redesign** può trovarsi a:
 - Definire o ridefinire le **tipologie** di documenti accolti nell'archivio
 - Definire o ridefinire i **template** di analisi del documento
 - Definire o ridefinire i **metadati** da estrarre dal documento
 - Definire o ridefinire i **pattern di estrazione** o di **validazione** di un metadato dal documento



TIPOLOGIE DI DOCUMENTI

I documenti si classificano in tipologie basandosi su

- Carattere
 - I documenti possono essere **scritti nativamente in forma digitale** oppure essere **scanditi**, quando sono scanditi possono essere **associati** o no ad un documento contenente il testo ottenuto dall'analisi della scansione mediante OCR;
- Provenienza
 - I documenti possono provenire da una fonte di tipo **esterno**, oppure essere prodotti da una applicazione di un ecosistema **interno**.
- Formato
 - I documenti possono essere in formato **pdf**, oppure in formato **jpeg**, o infine possono essere in formati intelleggibili standard come **html**.



TEMPLATE DI DOCUMENTI

Ad ogni **categoria** di documento, associata ai documenti stessi, vengono potenzialmente associati dei template che ci dicono **come è disposta l'informazione** all'interno del documento.

I template descrivono i documenti sulla base della loro organizzazione **logica** (contenuto e legame con i metadati) e **fisica** (disposizione del contenuto nel testo).

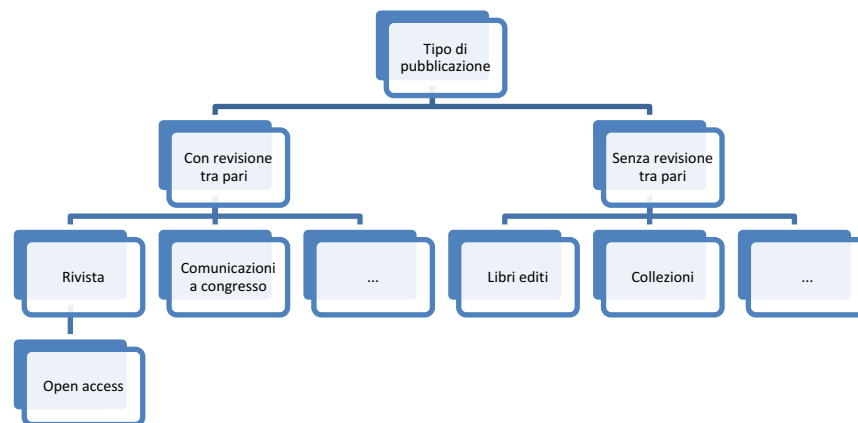
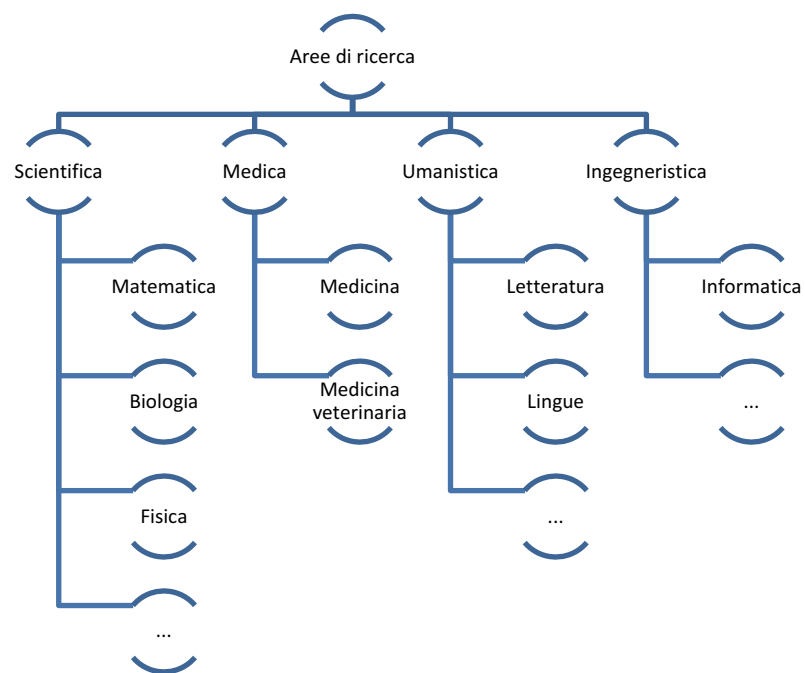


TASSONOMIE DI DOCUMENTI

Consideriamo un caso tipico universitario: la classificazione di documenti scientifici.

Ogni documento scientifico può essere inquadrato per categorie secondo uno schema che prevede tre organizzazioni:

- Per **categoria** di pubblicazione;
- Per **area di ricerca**;





TEMPLATE DOCUMENTALI

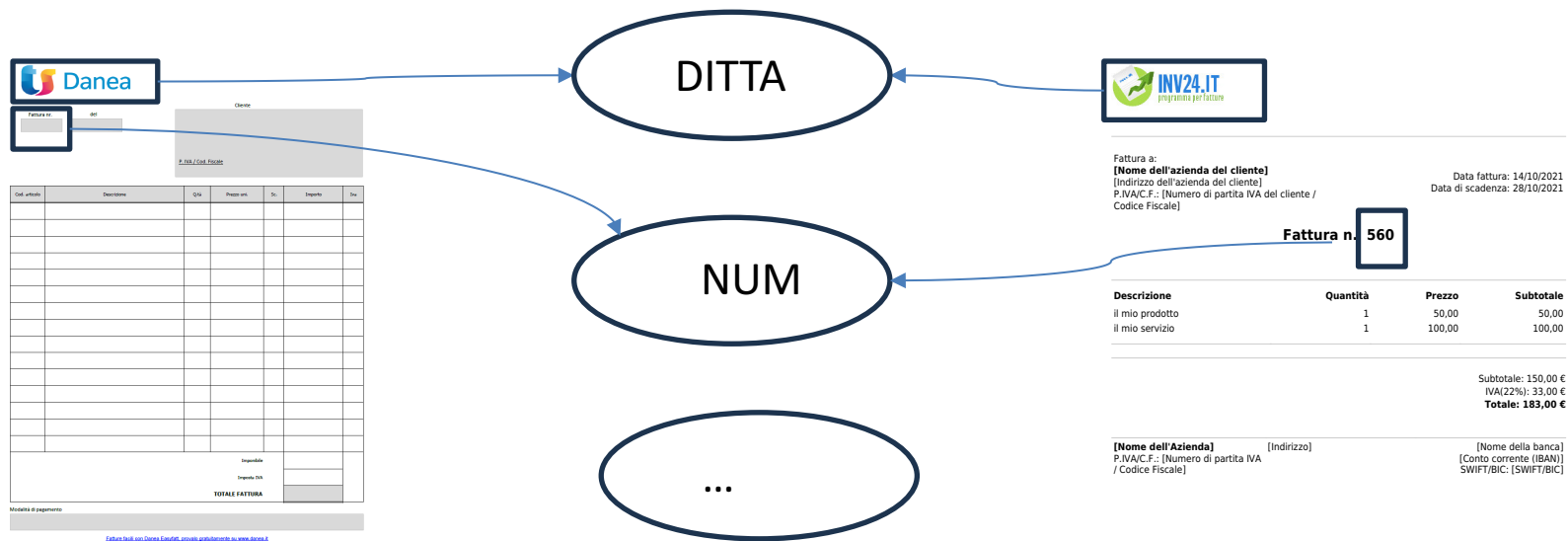
Nei documenti i contenuti e i metadati sono disposti in modo **organizzato**;

L'organizzazione dei dati, in molti casi, corrisponde a

- Parti fisiche (**aree**) del documento dove si trovano contenuti **che possono corrispondere** a dei metadati;
- Contenuti del documento che hanno una relazione **logica** o **fisica** con il metadato.



ESEMPI DI TEMPLATE: FATTURE





AUTOMAZIONE DI UN ARCHIVIO:

1. Riconoscimento automatico dei pattern di archiviazione/validazione
2. Generazione automatica dei metadati
3. Creazione di un sistema automatico di caricamento



FLUSSO DATI IN UN SISTEMA DI ARCHIVIAZIONE (AUTOMATICA)

1. Inserimento del documento in una cartella attiva
2. Triggering di classificazione del documento che ne riconosce pattern di categorizzazione nelle tassonomie associate
3. Associazione del documento ad uno o più template
4. Associazione del documento ai set di metadati che corrispondono a ciascun template
5. Riconoscimento dei pattern di identificazione dei singoli metadati per il template attivo
6. Inserimento dei metadati nel record del documento in archivio
7. Archiviazione effettiva



ALCUNI PRINCIPI DI DESIGN DEGLI ARCHIVI

- I **fattori di progettazione** vanno considerati tutti.
- Le **dinamiche di progettazione** vanno considerate tutte.
- Va adottato un **metodo di progettazione**.



FATTORI DI PROGETTAZIONE





GLI UTENTI DI UN ARCHIVIO

- Funzioni
 - Design (redesign)
 - Editing metadati
 - Inserimento in archivio
- Ruoli
 - Designer
 - Editor (potenzialmente per categoria)
 - Loader (potenzialmente per categoria)



DINAMICHE DI PROGETTAZIONE

Le domande che dobbiamo porci sono quindi quattro:

1. Perché il sistema di archiviazione **cambia**?
2. Quali aspetti rilevanti rispetto alla classificazione possono essere **condivisi** con altri archivi?
3. In quale modo un sistema di archiviazione **dipende** da aspetti di carattere legale?



METODI DI PROGETTAZIONE

Attualmente i pattern di design più importanti per gli archivi sono

1. Web-based inrerfacing (tipicamente in architettura SOAP/REST);
2. Cloud disposal, soprattutto Public cloud;
3. Sistemi di classificazione documentale standardizzati quali DUBLIN CORE.



DUBLIN CORE

Il Dublin Core è una iniziativa volta alla standardizzazione dei set di metadati utilizzati per l'archiviazione documentale, specificamente quelli che sono utilizzati in ambito **library management**

CLASSIFICATORI

- **Nome** – Etichetta assegnata al dato
- **Identificatore** – Identificativo univoco assegnato al dato
- **Versione** – Versione del dato
- **Registrazione di autorità** – Entità autorizzata a registrare il dato
- **Lingua** – Lingua nella quale il dato è indicato
- **Definizione** – Indicazione che rappresenta chiaramente il concetto e la natura essenziale del dato
- **Obbligatorietà** – Indica se il dato è richiesto sempre o solo in alcuni casi (contiene un valore)
- **Tipo di dato** – Indica la tipologia del dato che può essere rappresentato nel valore del dato stesso
- **Occorrenza massima** – Indica un limite alla ripetibilità del dato



DUBLIN CORE (CONTINUA)

ELEMENTI

Titolo

Creatore

Soggetto

Descrizione

Editore

**Autore di contributo
subordinato**

Data

Tipo

Formato

Identificatore

Fonte

Lingua

Relazione

Copertura

Gestione dei diritti



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

FINE PRIMA PARTE



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

SECONDA PARTE: INDICI, ONTOLOGIE E LINGUAGGIO



CLASSIFICARE DOCUMENTI

1. Compito automatizzabile che consiste nello stabilire quali siano, in un sistema di classi, le classi con la massima probabilità di contenere un certo documento dato in input
2. Concetto base: **distanza documentale**



METODI DI CLASSIFICAZIONE

- Metodi che mettono nella stessa classe documenti che sono sufficientemente simili
- Metodi che mettono in una classe documenti che assomigliano ad un documento master
- Metodi che inseriscono in una classe i documenti che contengono certe informazioni, o che hanno certe caratteristiche meramente associabili a metadati.



CHE COS'È LA DISTANZA TRA TESTI?

Questa domanda è difficile per vari motivi

- Il numero di differenze potrebbe non essere in alcun modo rilevante. Testi quasi identici dal punto di vista semantico potrebbero risultare del tutto differenti sul piano della misura delle differenze;
- Testi molto simili potrebbero essere molto diversi quando misurati con una o altra metodica di misura della distanza;
- Testi di lunghezza molto diversa ma che hanno molto in comune potrebbero essere molto dissimili se misurata la distanza in un modo troppo elementare.



HAMMING, LEVENSTHEIN, LA DISTANZA COSENO E IL LORO ACERRIMO NEMICO...L'INDICE TF-IDF

- La distanza di Hamming e di Levensthein sono misure basiche definite sul vocabolario (o sull'alfabeto) e non tengono conto della misura dimensionale dei testi.
- La distanza coseno lo fa.
- Ma in generale le parole di un vocabolario non sono l'insieme giusto su cui prendere queste misure, bisognerebbe lavorare sul lessico, ovvero sulle forme «base» delle parole.



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

DIGRESSIONE 1: LA MINIMUM EDIT DISTANCE



How similar are two strings?

Spell correction

The user typed “graffe”

Which is closest?

graf
graft
grail
giraffe

- Computational Biology

- Align two sequences of nucleotides

```
AGGCTATCACCTGACCTCCAGGCCGATGCCC  
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

- Resulting alignment:

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---  
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

- Also for Machine Translation, Information Extraction, Speech Recognition



Edit Distance

The minimum edit distance between two strings
Is the minimum number of editing operations

Insertion

Deletion

Substitution

Needed to transform one into the other



Minimum Edit Distance

Two strings and their **alignment**:

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N



Minimum Edit Distance

I N T E * N T I O N
| | | | | | | | |
* E X E C U T I O N
d s s i s

If each operation has cost of 1

Distance between these is 5

If substitutions cost 2 (Levenshtein)

Distance between them is 8



How to find the Min Edit Distance?

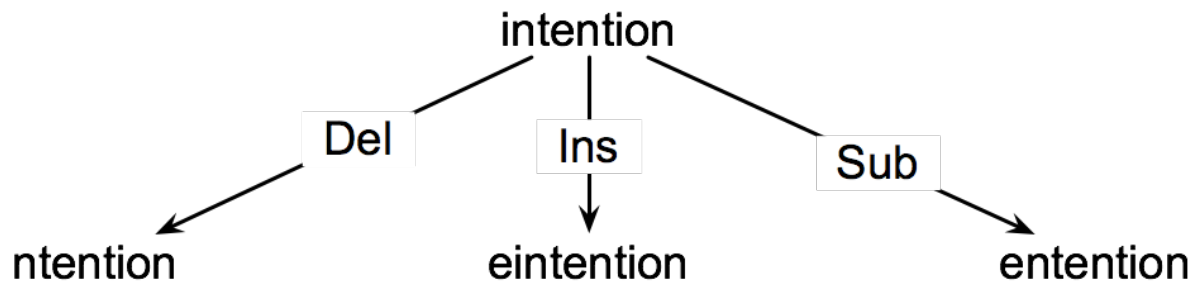
Searching for a path (sequence of edits) from the start string to the final string:

Initial state: the word we're transforming

Operators: insert, delete, substitute

Goal state: the word we're trying to get to

Path cost: what we want to minimize: the number of edits





Minimum Edit as Search

But the space of all edit sequences is huge!

We can't afford to navigate naïvely

Lots of distinct paths wind up at the same state.

We don't have to keep track of all of them

Just the shortest path to each of those revisited states.



Defining Min Edit Distance

For two strings

X of length n

Y of length m

We define $D(i,j)$

the edit distance between $X[1..i]$ and $Y[1..j]$

i.e., the first i characters of X and the first j characters of Y

The edit distance between X and Y is thus $D(n,m)$



Performance

Time:

$O(nm)$

Space:

$O(nm)$

Backtrace

$O(n+m)$



Weighted Edit Distance

Why would we add weights to the computation?

Spell Correction: some letters are more likely to be mistyped than others

Biology: certain kinds of deletions or insertions are more likely than others



Confusion matrix for spelling errors

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X	Y (correct)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	6	4	12	0	0	2	0	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**





UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

DIGRESSIONE 2: GLI N-GRAMMI



Regular expressions

A formal language for specifying text strings

How can we search for any of these?

woodchuck

woodchucks

Woodchuck

Woodchucks





Regular Expressions: Disjunctions

Letters inside square brackets []

	Pattern	Matches
Ranges [A	<code>[wW]oodchuck</code>	Woodchuck, woodchuck
	<code>[1234567890]</code>	Any digit

Pattern	Matches	
<code>[A-Z]</code>	An upper case letter	<u>D</u> renched Blossoms
<code>[a-z]</code>	A lower case letter	<u>m</u> y beans were impatient
<code>[0-9]</code>	A single digit	Chapter <u>1</u> : Down the Rabbit Hole



Regular Expressions: Negation in Disjunction

Negations [^Ss]

Carat means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	<u>I</u> have no exquisite reason"
[^e^]	Neither e nor ^	Look <u>h</u> ere
a^b	The pattern a carat b	Look up <u>a^b</u> now



Regular Expressions: More Disjunction

Woodchuck is another name for groundhog!
The pipe | for disjunction

Pattern	Matches
<code>groundhog woodchuck</code>	woodchuck
<code>yours mine</code>	yours
<code>a b c</code>	= <code>[abc]</code>
<code>[gG]roundhog [Ww]oodchuck</code>	Woodchuck



Regular Expressions: ? *+.



UNIVERSITÀ
di VERONA

Dipartimento
di INFORMATICA

Pattern	Matches	
<code>colou?r</code>	Optional previous char	<u>color</u> <u>colour</u>
<code>oo*h!</code>	0 or more of previous char	<u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u>
<code>o+h!</code>	1 or more of previous char	<u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u>
<code>baa+</code>		<u>baa</u> <u>baaa</u> <u>baaaa</u> <u>baaaaa</u>
<code>beg.n</code>		<u>begin</u> <u>begun</u> <u>begun</u> <u>beg3n</u>



Stephen C Kleene

Kleene *, Kleene +



Regular Expressions: Anchors [^] \$

Pattern	Matches
[^] [A-Z]	<u>P</u> alo Alto
[^] [[^] A-Za-z]	<u>1</u> "Hello"
\. \$	The end <u>.</u>
. \$	The end <u>?</u> The end <u>!</u>



Example

Find me all instances of the word “the” in a text.

the

Misses capitalized examples

[tT]he

Incorrectly returns other or theology

[^a-zA-Z][tT]he[^a-zA-Z]



Errors

The process we just went through was based on fixing two kinds of errors:

1. Matching strings that we should not have matched
(there, then, other)

False positives (Type I errors)

2. Not matching things that we should have matched
(The)

False negatives (Type II errors)



Errors cont.

In NLP we are always dealing with these kinds of errors. Reducing the error rate for an application often involves two antagonistic efforts:

- Increasing accuracy or precision (minimizing false positives)
- Increasing coverage or recall (minimizing false negatives).



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

DIGRESSIONE 3: IL BAG OF WORDS



How many words in a sentence?

"I do uh main- mainly business data processing"

Fragments, filled pauses

"Seuss's **cat** in the hat is different from other **cats!**"

Lemma: same stem, part of speech, rough word sense

cat and **cats** = same lemma

Wordform: the full inflected surface form

cat and **cats** = different wordforms



How many words in a sentence?

they lay back on the San Francisco grass and looked at the stars
and their

Type: an element of the vocabulary.

Token: an instance of that type in running text.

How many?

15 tokens (or 14)

13 types (or 12) (or 11?)



How many words in a corpus?

N = number of tokens

V = vocabulary = set of types, **|V|** is size of vocabulary

Heaps Law = Herdan's Law = $|V| = kN^\beta$ where often $.67 < \beta < .75$

i.e., vocabulary size grows with $>$ square root of the number of word tokens

	Tokens = N	Types = V
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
COCA	440 million	2 million
Google N-grams	1 trillion	13+ million



Corpora

Words don't appear out of nowhere!

A text is produced by

- a specific writer(s),
- at a specific time,
- in a specific variety,
- of a specific language,
- for a specific function.



Corpora vary along dimension like

Language: 7097 languages in the world

Variety, like African American Language varieties.

AAE Twitter posts might include forms like "*iont*" (*I don't*)

Code switching, e.g., Spanish/English, Hindi/English:

S/E: Por primera vez veo a @username actually being hateful! It was beautiful:)

[For the first time I get to see @username actually being hateful! it was beautiful:]

H/E: dost tha or ra- hega ... dont worry ... but dherya rakhe

["he was and will remain a friend ... don't worry ... but have faith"]

Genre: newswire, fiction, scientific articles, Wikipedia

Author Demographics: writer's age, gender, ethnicity, SES



Text Normalization

Every NLP task requires text normalization:

1. Tokenizing (segmenting) words
2. Normalizing word formats
3. Segmenting sentences



Space-based tokenization

A very simple way to tokenize

For languages that use space characters between words

Arabic, Cyrillic, Greek, Latin, etc., based writing systems

Segment off a token between instances of spaces

Unix tools for space-based tokenization

The "tr" command

Inspired by Ken Church's UNIX for Poets

Given a text file, output the word tokens and their frequencies



Lemmatization is done by Morphological Parsing

Morphemes:

The small meaningful units that make up words

Stems: The core meaning-bearing units

Affixes: Parts that adhere to stems, often with grammatical functions

Morphological Parsers:

Parse *cats* into two morphemes *cat* and *s*

Parse Spanish *amaren* ('if in the future they would love') into morpheme *amar* 'to love', and the morphological features *3PL* and *future subjunctive*.



Stemming

Reduce terms to stems, chopping off affixes crudely

This was not the map we found in Billy Bones's chest, but an accurate copy, complete in all things-names and heights and soundings-with the single exception of the red crosses and the written notes.



This was not the map we found in Billy Bones's chest but an accurate copy complete in all things names and heights and soundings with the single exception of the red crosses and the written notes.



Dealing with complex morphology is necessary for many languages

e.g., the Turkish word:

Uygarlastiramadiklarimizdanmissinizcasina

‘(behaving) as if you are among those whom we could not civilize’

Uygar ‘civilized’ + **las** ‘become’

+ **tir** ‘cause’ + **ama** ‘not able’

+ **dik** ‘past’ + **lar** ‘plural’

+ **imiz** ‘p1pl’ + **dan** ‘abl’

+ **mis** ‘past’ + **siniz** ‘2pl’ + **casina** ‘as if’



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

DIGRESSIONE 4: IL FANTASTICO MONDO DI AMELIE



Probabilistic Language Modeling

Goal: compute the probability of a sentence or sequence of words:

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \dots w_n)$$

Related task: probability of an upcoming word:

$$P(w_5 | w_1, w_2, w_3, w_4)$$

A model that computes either of these:

$P(W)$ or $P(w_n | w_1, w_2 \dots w_{n-1})$ is called a **language model**.

Better: **the grammar** But **language model** or **LM** is standard



How to compute $P(W)$

How to compute this joint probability:

$P(\text{its, water, is, so, transparent, that})$

Intuition: let's rely on the Chain Rule of Probability



Reminder: The Chain Rule

Recall the definition of conditional probabilities

$$p(B|A) = P(A,B)/P(A) \quad \text{Rewriting: } P(A,B) = P(A)P(B|A)$$

More variables:

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$

The Chain Rule in General

$$P(x_1, x_2, x_3, \dots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \dots P(x_n|x_1, \dots, x_{n-1})$$



The Chain Rule applied to compute joint probability of words in sentence

$P(\text{"its water is so transparent"}) =$
 $P(\text{its}) \times P(\text{water} | \text{its}) \times P(\text{is} | \text{its water})$
 $\times P(\text{so} | \text{its water is}) \times P(\text{transparent} | \text{its water is so})$

$$P(w_1 w_2 \dots w_n) = \prod_i P(w_i | w_1 w_2 \dots w_{i-1})$$



Andrei Markov

Markov Assumption

Simplifying assumption:

$$P(\text{the } l \text{ its water is so transparent that}) \approx P(\text{the } l \text{ that})$$

Or maybe

$$P(\text{the } l \text{ its water is so transparent that}) \approx P(\text{the } l \text{ transparent that})$$



Markov Assumption

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i | w_{i-k} \dots w_{i-1})$$

In other words, we approximate each component in the product

$$P(w_i | w_1 w_2 \dots w_{i-1}) \approx P(w_i | w_{i-k} \dots w_{i-1})$$



Simplest case: Unigram model

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i)$$

Some automatically generated sentences from a unigram model

fifth, an, of, futures, the, an, incorporated, a,
a, the, inflation, most, dollars, quarter, in, is,
mass

thrift, did, eighty, said, hard, 'm, july, bullish

that, or, limited, the



Bigram model

- Condition on the previous word:

$$P(w_i | w_1 w_2 \dots w_{i-1}) \approx P(w_i | w_{i-1})$$

texaco, rose, one, in, this, issue, is, pursuing, growth, in,
a, boiler, house, said, mr., gurria, mexico, 's, motion,
control, proposal, without, permission, from, five, hundred,
fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november



Estimating bigram probabilities

The Maximum Likelihood Estimate

$$P(w_i | w_{i-1}) = \frac{\mathit{count}(w_{i-1}, w_i)}{\mathit{count}(w_{i-1})}$$

$$P(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$



Berkeley Restaurant Project sentences

Can you tell me about any good cantonese restaurants close by
mid priced thai food is what i'm looking for?

Tell me about chez panisse...

Can you give me a listing of the kinds of food that are available
i'm looking for a good place to eat breakfast?

When is caffe venezia open during the day?



Raw bigram counts

Out of 9222 sentences

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0



Raw bigram probabilities

Normalize by unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Result:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0



Language Modeling Toolkits

SRILM

<http://www.speech.sri.com/projects/srilm/>

KenLM

<https://kheafield.com/code/kenlm/>



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

DIGRESSIONE 5: Tf-IDf



Text Classification: definition

Input:

a document d

a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$

Output: a predicted class $c \in C$



Classification Methods: Hand-coded rules

Rules based on combinations of words or other features

spam: black-list-address OR (“dollars” AND “you have been selected”)

Accuracy can be high

If rules carefully refined by expert

But building and maintaining these rules is expensive



Classification Methods: Supervised Machine Learning

Input:

a document d

a fixed set of classes $C = \{c_1, c_2, \dots, c_j\}$

A training set of m hand-labeled documents

$(d_1, c_1), \dots, (d_m, c_m)$

Output:

a learned classifier $\gamma: d \rightarrow c$



Classification Methods: Supervised Machine Learning

Any kind of classifier

Naïve Bayes

Logistic regression

Neural networks

k-Nearest Neighbors

...



Naive Bayes Intuition

Simple ("naive") classification method based on Bayes rule

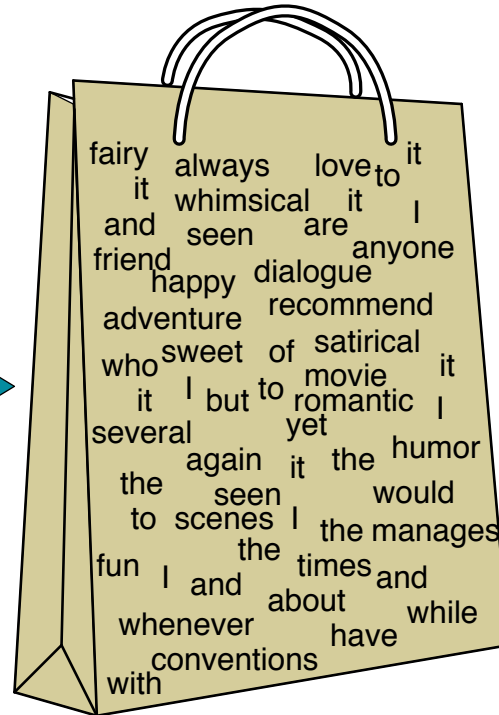
Relies on very simple representation of document

Bag of words



The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



The bag of words representation

$Y =$

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

$= C$






Bayes' Rule Applied to Documents and Classes

- For a document d and a class c

$$P(c | d) = \frac{P(d | c)P(c)}{P(d)}$$



Naive Bayes Classifier (I)

$$C_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

Dropping the denominator



Naive Bayes Classifier (II)

"Likelihood"

"Prior"

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c)P(c)$$

Document d
represented as
features x1..xn



Stop words

Often systems ignore **stop words**

Stop words: very frequent words like *the* and *a*.

Sort the vocabulary by word frequency in training set

Call the top 10 or 50 words the **stopword list**.

Remove all stop words from both training and test sets As if they were never there!



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

DIGRESSIONE 6: VALUTAZIONE DEI MODELLI



The 2-by-2 confusion matrix

gold standard labels

		gold positive	gold negative	
<i>system output labels</i>	system positive	true positive	false positive	precision = $\frac{tp}{tp+fp}$
	system negative	false negative	true negative	
		recall = $\frac{tp}{tp+fn}$		accuracy = $\frac{tp+tn}{tp+fp+tn+fn}$



Evaluation: Accuracy

Why don't we use **accuracy** as our metric?

Imagine we saw 1 million tweets

100 of them talked about Delicious Pie Co.

999,900 talked about something else

We could build a dumb classifier that just labels every tweet "not about pie"

It would get 99.99% accuracy!!! Wow!!!!

But useless! Doesn't return the comments we are looking for!

That's why we use **precision** and **recall** instead



Evaluation: Precision

% of items the system detected (i.e., items the system labeled as positive) that are in fact positive (according to the human gold labels)

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$



Evaluation: Recall

% of items actually present in the input that were correctly identified by the system.

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$



A combined measure: F

F measure: a single number that combines P and R :

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

We almost always use balanced F_1 (i.e., $\beta = 1$)

$$F_1 = \frac{2PR}{P + R}$$



Development Test Sets ("Devsets") and Cross-validation

Training set

Development Test Set

Test Set

Train on training set, tune on devset, report on testset

This avoids overfitting ('tuning to the test set')

More conservative estimate of performance

But paradox: want as much data as possible for training, and as much for dev; how to split?



Confusion Matrix for 3-class classification

gold labels

		urgent	normal	spam	
<i>system output</i>	urgent	8	10	1	precision_u = $\frac{8}{8+10+1}$
	normal	5	60	50	precision_n = $\frac{60}{5+60+50}$
	spam	3	30	200	precision_s = $\frac{200}{3+30+200}$
		recall_u = $\frac{8}{8+5+3}$	recall_n = $\frac{60}{10+60+30}$	recall_s = $\frac{200}{1+50+200}$	



Macroaveraging and Microaveraging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = .73$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = .60$$



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

ATTERRAGGIO: I LARGE LANGUAGE MODELS



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

FINE SECONDA PARTE



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

TERZA PARTE: L'AMICO IMMAGINARIO



SPUNTA IL CONVITATO DI PIETRA: CHATGPT

Q: Can you prepare a presentation in powerpoint about ChatGPT?

A: Certainly! I can provide you with an outline for a PowerPoint presentation about ChatGPT. Feel free to customize it and add your own content and design elements.



UNN EFFETTO QUASI MAGICO

Slide 1:

- Title Slide
- Include the presentation title, your name, and any relevant information

Slide 2:

- Introduction
- Briefly introduce the topic of the presentation
- Explain the importance of conversational AI and its applications

Slide 3:

- What is ChatGPT?
- Provide an overview of ChatGPT
- Explain that ChatGPT is a language model developed by OpenAI
- Mention that it is based on the GPT-3.5 architecture



Slide 4:

- How does ChatGPT work?
- Explain the underlying technology of ChatGPT
- Mention that it is based on deep learning and natural language processing (NLP) techniques
- Describe how it processes and generates human-like responses

Slide 5:

- Key Features of ChatGPT
- Discuss the important features of ChatGPT
- Emphasize its ability to understand and generate human-like text
- Mention its support for multiple languages

Slide 6:

- Applications of ChatGPT
- Highlight various applications of ChatGPT
- Include examples such as customer support, language translation, content creation, and tutoring



Slide 7:

- Benefits of ChatGPT
- Discuss the advantages of using ChatGPT
- Highlight increased efficiency, scalability, and cost-effectiveness
- Mention its potential to improve customer experience and user engagement

Slide 8:

- Limitations and Challenges
- Discuss the limitations and challenges of ChatGPT
- Mention issues such as bias, lack of context, and potential ethical concerns
- Emphasize the need for careful deployment and human oversight

Slide 9:

- Future Developments
- Discuss the future of ChatGPT and conversational AI
- Mention ongoing research and development efforts
- Highlight the potential for improved models and expanded capabilities



Slide 10:

- Conclusion
- Summarize the main points discussed in the presentation
- Reiterate the significance of ChatGPT in the field of conversational AI

Slide 11:

- Questions and Answers
- Encourage the audience to ask questions and engage in a discussion

Slide 12:

- Thank You
- Thank the audience for their attention
- Provide contact information if necessary

This is just a basic outline to get you started. You can add more slides, include visuals, and expand on each topic as needed. Good luck with your presentation!



Q: How can be ChatGPT used in managing Document Archives?

Slide 1:

- Title Slide
- Include the presentation title, your name, and any relevant information

Slide 2:

- Introduction
- Briefly introduce the topic of managing document archives
- Highlight the challenges and complexities of handling large amounts of documents

Slide 3:

- Document Archive Management
- Explain the importance of effective document archive management
- Highlight the need for organization, searchability, and accessibility

Slide 4:

- Introducing ChatGPT in Document Archives
- Explain how ChatGPT can be utilized in document archive management
- Mention its natural language processing capabilities and conversational interface



Slide 5:

- Intelligent Search and Retrieval
- Discuss how ChatGPT can enhance search and retrieval functionalities
- Explain how users can interact with ChatGPT to find specific documents or information

Slide 6:

- Contextual Document Summarization
- Highlight ChatGPT's ability to generate contextual summaries
- Explain how it can analyze and summarize lengthy documents for quick understanding

Slide 7:

- Automatic Categorization and Tagging
- Discuss how ChatGPT can assist in automatically categorizing and tagging documents
- Explain its ability to understand document content and assign appropriate labels

Slide 8:

- Language Translation and Localization
- Explain how ChatGPT can aid in translating documents between different languages
- Discuss its potential for providing localized versions of documents



Slide 9:

- Workflow Automation and Assistance
- Highlight ChatGPT's potential in automating document-related workflows
- Explain how it can assist users in generating documents or performing repetitive tasks

Slide 10:

- Collaboration and Knowledge Sharing
- Discuss how ChatGPT can facilitate collaboration and knowledge sharing within document archives
- Explain its ability to answer questions, provide insights, and suggest related documents

Slide 11:

- Benefits of Using ChatGPT in Document Archives
- Summarize the key benefits of integrating ChatGPT in document archive management
- Emphasize improved search efficiency, enhanced document understanding, and increased productivity



Slide 12:

- Limitations and Considerations
- Discuss the limitations and considerations when using ChatGPT in document archives
- Mention potential issues such as accuracy, bias, and security concerns
- Emphasize the need for human oversight and continuous improvement

Slide 13:

- Future Possibilities
- Discuss future possibilities and advancements in using ChatGPT for document archive management
- Mention ongoing research and development efforts in the field

Slide 14:

- Conclusion
- Summarize the main points discussed in the presentation
- Highlight the potential of ChatGPT in transforming document archive management

Slide 15:

- Questions and Answers
- Encourage the audience to ask questions and engage in a discussion



Slide 16:

- Thank You
- Thank the audience for their attention
- Provide contact information if necessary

Feel free to expand on each slide with more details, examples, or visuals as needed. Good luck with your presentation!



Q: Let the slides apart. Can you tell me how ChaGPT can be used in the practice of document classification and retrieval?

A: Certainly! ChatGPT can be used in the practice of document classification and retrieval in the following ways:

1. **Intelligent Document Search and Retrieval:** ChatGPT can enhance traditional search and retrieval functionalities by providing a conversational interface. Users can interact with ChatGPT using natural language queries, making the process more intuitive and user-friendly. ChatGPT can understand the query context and provide relevant document suggestions or directly retrieve specific documents based on the user's request.
2. **Contextual Document Summarization:** ChatGPT's language processing capabilities enable it to analyze and understand the content of documents. It can generate contextual summaries, extracting key information and capturing the essence of lengthy documents. This feature helps users quickly understand the document's content and relevance before deciding to access the full document.



3. **Automatic Categorization and Tagging:** ChatGPT can assist in automatically categorizing and tagging documents. By training ChatGPT on existing document archives or using supervised learning techniques, it can learn to recognize patterns and assign appropriate labels or tags to documents. This automation streamlines the categorization process, making it easier to organize and retrieve documents based on specific categories or tags.
4. **Language Translation and Localization:** ChatGPT's multilingual capabilities make it useful for document classification and retrieval across different languages. It can aid in translating documents between languages, enabling users to access documents in their preferred language. Furthermore, ChatGPT can assist in providing localized versions of documents, adapting content to specific regions or cultural contexts.
5. **Workflow Automation and Assistance:** ChatGPT can automate document-related workflows and assist users in performing tasks. It can generate standard document templates, complete form fields, or extract specific information from documents. This automation improves efficiency and accuracy, freeing up time for users to focus on more critical tasks.



6. Collaboration and Knowledge Sharing: ChatGPT's conversational nature allows it to facilitate collaboration and knowledge sharing within document archives. Users can ask questions, seek clarifications, or request insights from ChatGPT. It can provide relevant information, suggest related documents, or help identify experts within the organization who have knowledge on specific topics, thereby fostering a collaborative environment.
7. By leveraging ChatGPT's language understanding and conversational capabilities, document classification and retrieval processes can be enhanced, providing users with more efficient and intuitive ways to manage and access their document archives.



Q: Can you make an example of an interaction between an imaginary user and ChatGPT in the retrieval of a specific set of documents in a document archive?

User: Hello, ChatGPT! I'm looking for documents related to project management methodologies. Can you help me retrieve them from our document archive?

ChatGPT: Of course! I'd be happy to assist you. Could you please provide me with any specific keywords or additional details that might help narrow down the search?

User: Sure! I'm particularly interested in documents that cover agile project management and scrum methodology.

ChatGPT: Great! I'll search our document archive for documents related to agile project management and scrum methodology. Please give me a moment.

[ChatGPT performs a search in the document archive]



ChatGPT: I have found several documents that match your search criteria. Here are the top three results:

- 1."Introduction to Agile Project Management": This document provides a comprehensive overview of agile project management principles, including the scrum methodology. It covers key concepts, roles, and processes involved in agile project management.
- 2."Scrum Master's Guide: Best Practices and Implementation Tips": This document focuses specifically on the scrum methodology and provides guidance for scrum masters. It offers practical tips, case studies, and real-world examples to help teams implement scrum effectively.
- 3."Comparing Agile Project Management Frameworks": This document compares different agile project management frameworks, including scrum, Kanban, and Lean. It highlights the strengths and weaknesses of each framework, helping you make informed decisions in selecting the right approach for your projects.

User: Excellent! The first document sounds like what I need. Can you provide a summary of it?



ChatGPT: Certainly! Here's a summary of the "Introduction to Agile Project Management" document: This document offers a comprehensive overview of agile project management, focusing on the scrum methodology. It explains the key principles behind agile project management and provides insights into how cross-functional teams work together in iterative cycles. It covers important concepts such as user stories, sprint planning, and daily stand-ups. Additionally, it explores the roles of the product owner, scrum master, and development team, emphasizing their responsibilities and collaboration. Overall, this document serves as a valuable resource for understanding the foundations of agile project management and scrum methodology.

User: That's perfect! Thank you for your help, ChatGPT.

ChatGPT: You're welcome! If you need any further assistance or have more questions, feel free to ask.



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

UN PO' DI PROVE?



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

FINE
