COMP5423 Natural Language Processing

Text Classification and Ranking

Outline

- **Learning Objectives**
  - Text Classification
    - Classification Models
    - Linguistic Features
  - Text Ranking
    - Relevance and Cosine Similarity
    - Bag-of-Word Retrieval vs. Dense Retrieval

Text Classification

- **Text Classification Tasks**
  - Binary Classification
    - Sentiment Classification: Determines whether the sentiment orientation that a writer expresses towards some object is positive or negative.
    - Email Spam Detection: Detects whether an email is spam or not.
  - Multi-Class Classification
    - News Categorization: Identifies the topic that a news talks about, such as business, technology, entertainment, sports, science and health, etc.

Text Classification

- **Sentiment Classification**

A Document

↓

Binary Sentiment Classification

↓

Positive or Negative

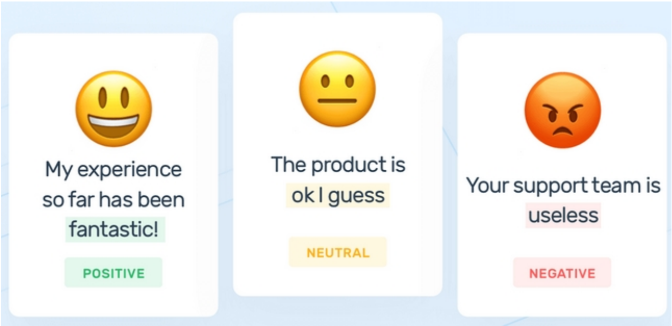
A Sentence

↓

Three-Class Sentiment Classification

↓

Positive, Negative or Neutral



A sentence may express no opinion. No opinion is usually regarded as neutral.

Text Classification

- **Sentiment Classification**

Information Extraction

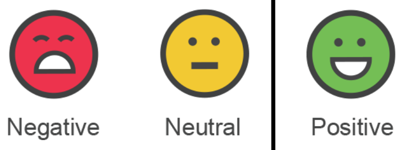
(Named) Entity Extraction

Co-Reference Resolution

Opinion Holder

Opinion Target

Sentiment



Text Classification

Sentiment Analysis and Opinion Mining

Text Classification

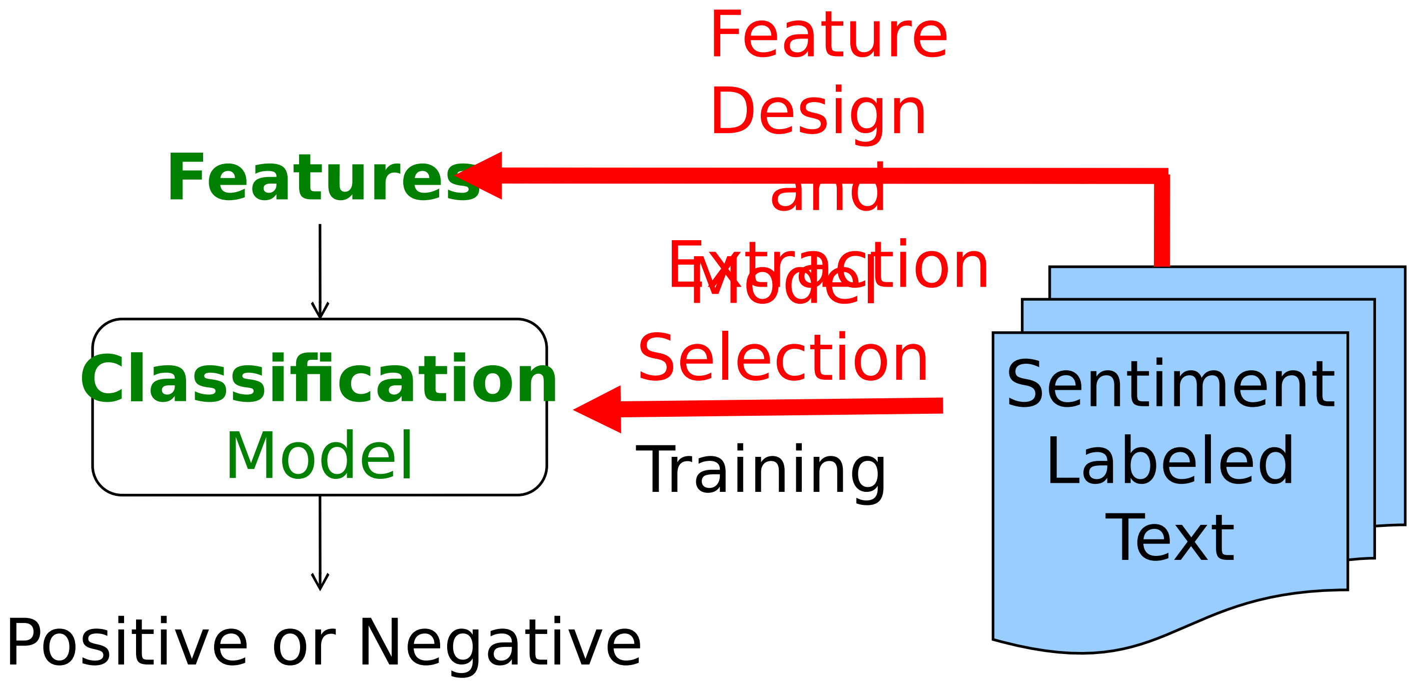
- **Sentiment Classification**
  - Supervised Classification: Training Corpus

+	It is a very amazing product.
+	Just received this camera two days ago and already love the features it has.
+	By cocking the shutter to the halfway position and getting the settings ready to shoot, I was able to produce excellent stopaction photos.
-	It feels slow to focus, and unbearably slow to shoot.
-	The adobe camera raw plug-in shows once again that hardware is miles ahead of software.
	……

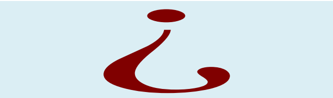
Sample Training Data

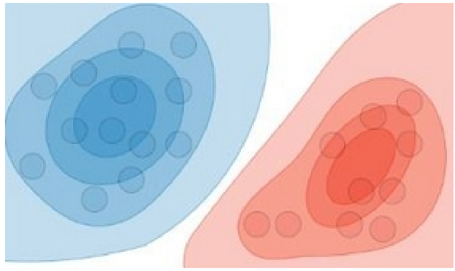
Text Classification

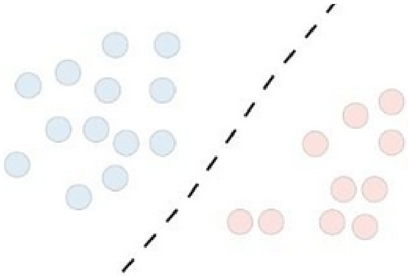
- **Sentiment Classification**
  - Supervised Classification: Training Corpus



Text Classification

- **Sentiment Classification**
  - Classification Models: Naive Bayes (NB), Logistic Regression (LR) (or Maximum Entropy (ME)), Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN), BERT, etc.
    - Generative Model (): Naïve Bayes Classifier (see Speech and Language Processing Chapter 4)
    - Discriminative Model (): Logistic Regression Classifier (see Speech and Language Processing Chapter 5)





Text Classification

- **Sentiment Classification**
  - Classification Features: Like other supervised machine learning applications, the key for sentiment classification is the engineering of a set of effective features.
  - Surface Text Features
    - (TF or TF-IDF Weighted) Word *N*-Grams, such as Word Uni-gram Features, Bi-Gram Features and etc.

Text Classification

It is a very amazing product.

6 Word Uni-gram Features (w/o Stop Word Removal and Stemming) Extracted

a	1
……	……
amazing	1
……	……
is	1
……	……
it	1
……	……
product	1
……	……
very	1
……	……

Uni-gram Feature Space

Text Classification

It is a very amazing product.

5 Word Bi-grams Features (w/o Stop Word Removal and Stemming) Extracted

......	......
a very	1
......	......
amazing product	1
......	......
is a	1
......	......
it is	1
......	......
very amazing	1
......	......

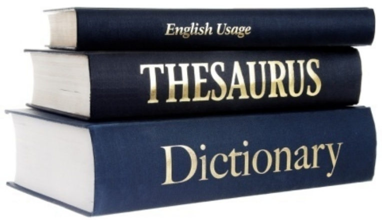
Bi-gram Feature Space

Text Classification

- **Sentiment Classification**
  - Popular Sentiment Lexicons
    - General Inquirer (GI)
    - LIWC
    - The Opinion Lexicon from HU and LIU
    - MPQA Subjectivity Lexicon
  - Lexicon-based Features
    - Sentiment Words and Phrases (from Sentiment Lexicon/Dictionary)

Text Classification

- **Sentiment Classification**



POSITIVE: good, wonderful, amazing, brilliant, perfect, beautiful, enjoy, love, favor, …

NEGATIVE: bad, poor, terrible, depressing, poorly, annoying, boring, sadly, bothered, …

It is a very amazing product.

1 Sentiment Word Feature Extracted

$$\begin{pmatrix} \ldots\ldots \\ \ldots\ldots \\ \text{amazing} \\ \ldots\ldots \\ \ldots\ldots \end{pmatrix} \quad \begin{pmatrix} \ldots\ldots \\ \ldots\ldots \\ 1 \\ \ldots\ldots \\ \ldots\ldots \end{pmatrix}$$

Lexicon Word Feature Space

Text Classification

- **Sentiment Classification**
  - Linguistic Features
    - Part-of-Speech (POS) Tags and their *N*-Gram Features (POS *N*-Grams)

It is a very amazing product.

POS Tagging

It/PRP is/VBZ a/DT very/RB amazing/JJ product/NN ./.

Text Classification

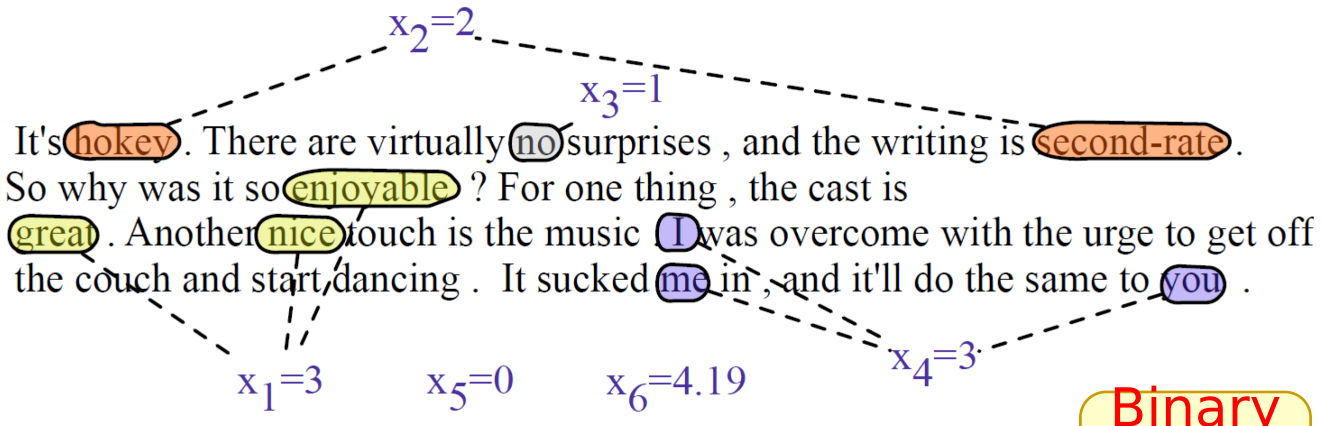
- **Sentiment Classification**
  - Linguistic Features

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	”	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

- Thumbs up? Sentiment Classification using Machine Learning Techniques

Text Classification

- **Sentiment Classification**
  - Other Statistic and Specially Designed Features



Binary Feature

Integer Feature

Real Value Feature

$$
\begin{array}{ll}
x_1 & \text{count(positive lexicon)} \in \text{doc)} \\
x_2 & \text{count(negative lexicon)} \in \text{doc)} \\
x_3 & \begin{cases} 1 & \text{if “no”} \in \text{doc} \\ 0 & \text{otherwise} \end{cases} \\
x_4 & \text{count(1st and 2nd pronouns} \in \text{doc)} \\
x_5 & \begin{cases} 1 & \text{if “!”} \in \text{doc} \\ 0 & \text{otherwise} \end{cases} \\
x_6 & \log(\text{word count of doc})
\end{array}
$$

Text Classification

- **Email Spam Detection**
  - Specially Designed Linguistic Features
    - Email subject line is all capital letters.
    - Email subject line contains “online pharmaceutical”.
    - Contains phrases of urgency like “urgent reply”.
    - Claims you can be removed from the list.
    - HTML has unbalanced “head” tags.
    - ….

Text Classification

- **Period Classification (End of Sentence or Not)**
  - Specially Designed Linguistic Features

$$x_1 = \begin{cases} 1 & \text{if ``}Case(w_i) = \text{Lower''} \\ 0 & \text{otherwise} \end{cases}$$

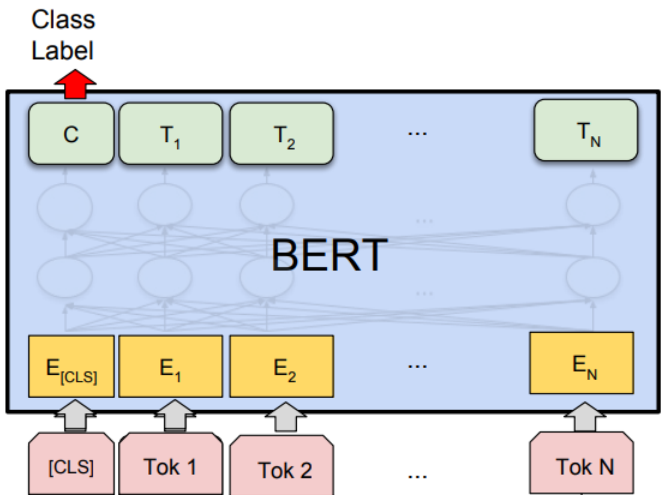
$$x_2 = \begin{cases} 1 & \text{if ``}w_i \in \text{AcronymDict''} \\ 0 & \text{otherwise} \end{cases}$$

Composite (Combination) Feature

$$x_3 = \begin{cases} 1 & \text{if ``}w_i = \text{St. \& } Case(w_{i-1}) = \text{Cap''} \\ 0 & \text{otherwise} \end{cases}$$

Text Classification

- **Representation of Input Text**
  - Feature Engineering: Features are generally designed by examining the training data set with an eye to linguistic intuitions and the linguistic literature on the domain.
    - In statistical natural language processing, feature design or feature selection is very important.
  - Representation Learning: In order to avoid extensive human effort of feature design, recent research in neural natural language processing has focused on representation learning.



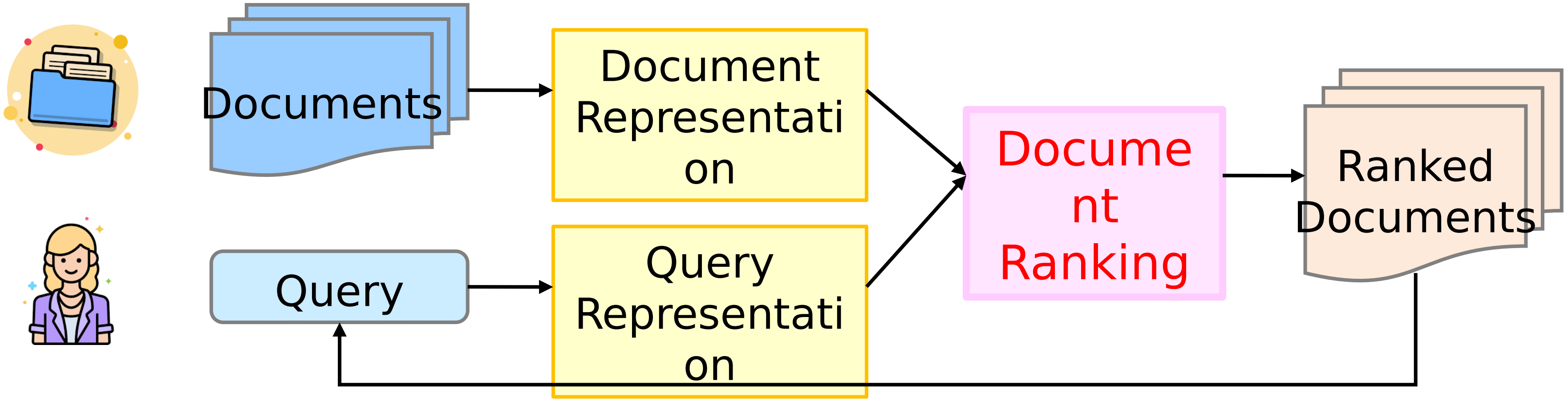
Bidirectional Encoder Representations from Transformers (BERT)

Text Classification

- **Chinese Word Segmentation via Classification**
  - Homework: Chinese word segmentation can be cast to a binary or multi-class classification problem. Do you have any idea how to apply a typical classification model (such as SVM, Naïve Bayes) to segment a given Chinese sentence into a sequence of words?

Text Ranking

- **Text Similarity (in Information Retrieval)**
  - Queries are treated as very short documents.



  - An information retrieval model needs a way to calculate the similarity between a query vector and a document vector as a measure of the relevance score of the document for that query (denoted by *sim*(*d*, *q*) or *score*(*d*, *q*)).

Text Ranking

- **Text Similarity**

  Assume and , where d_i and q_i represent the weight of the ith term in document d and query q respectively.

  - Euclidean Length of a Vector

  - Dot Product $$\vec{d} \cdot \vec{q} = d_1 q_1 + d_2 q_2 + \ldots + d_n q_n = \sum_{i=1}^{n} d_i q_i$$

Text Ranking

- **Text Similarity**
  - Cosine Similarity

Numerato
r: Dot
Product

¿ $$\frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

Euclidean
Normalizati
on

¿ $$\frac{\vec{d}\cdot\vec{q}}{|\vec{d}||\vec{q}|}$$

Denominator:
Euclidean
Lengths

Text Ranking

- **Text Similarity**

Document Set	*d*1	*d*2	*d*3	*d*4
*t*1	1	0	0	1
*t*2	0	1	0	1
*t*3	0	1	1	1
*t*4	0	1	1	0
*t*5	1	1	0	1

Query	*q*
*t*1	1
*t*2	0
*t*3	0
*t*4	0
*t*5	1

Text Ranking

[1] TF-IDF Weighed Vector Representation

	*d*1	*d*2	*d*3	*d*4
*t*1	1 log(4/2)=0.30	0	0	1 log(4/2)=0.30
*t*2	0	1 log(4/2)=0.30	0	1 log(4/2)=0.30
*t*3	0	1 log(4/3)=0.12	1 log(4/3)=0.12	1 log(4/3)=0.12
		1 log(4/2)=0.30	1 log(4/2)=0.30	0
		1 log(4/3)=0.12		1 log(4/3)=0.12

	q
*t*1	1 log(4/2)=0.30
*t*2	0
*t*3	0
*t*4	0
*t*5	1 log(4/3)=0.12

$$d_1=\begin{pmatrix}0.30\\0\\0\\0\\0.12\end{pmatrix}\quad d_2=\begin{pmatrix}0\\0.30\\0.12\\0.30\\0.12\end{pmatrix}\quad d_3=\begin{pmatrix}0\\0\\0.12\\0.30\\0\end{pmatrix}\quad d_4=\begin{pmatrix}0.30\\0.30\\0.12\\0\\0.12\end{pmatrix}\quad q=\begin{pmatrix}0.30\\0\\0\\0\\0.12\end{pmatrix}$$

Text Ranking

[2] Cosine Similarity

$$sim(\vec{d}_1,\vec{q})=\frac{0.3\times0.3+0.12\times0.12}{\sqrt{0.3^2+0.12^2}\sqrt{0.3^2+0.12^2}}=1 \qquad sim(\vec{d}_3,\vec{q})=\frac{0}{\sqrt{0.3^2+0.12^2}\sqrt{0.3^2+0.12^2}}=0$$

$$sim(\vec{d}_2,\vec{q})=\frac{0.12\times0.12}{\sqrt{2(0.3\text{¿¿}2+0.12^2)}\sqrt{0.3^2+0.12^2}=0.098\text{¿}} \qquad sim(\vec{d}_4,\vec{q})=\frac{0.3\times0.3+0.12\times0.12}{\sqrt{2(0.3\text{¿¿}2+0.12^2)}\sqrt{0.3^2+0.12^2}=0.707\text{¿}}$$

	d1	*d2*	*d3*	*d4*
similarity	1	0.098	0	0.707

- Question: Can you tell the scores for d1 and d3 without performing any calculation?

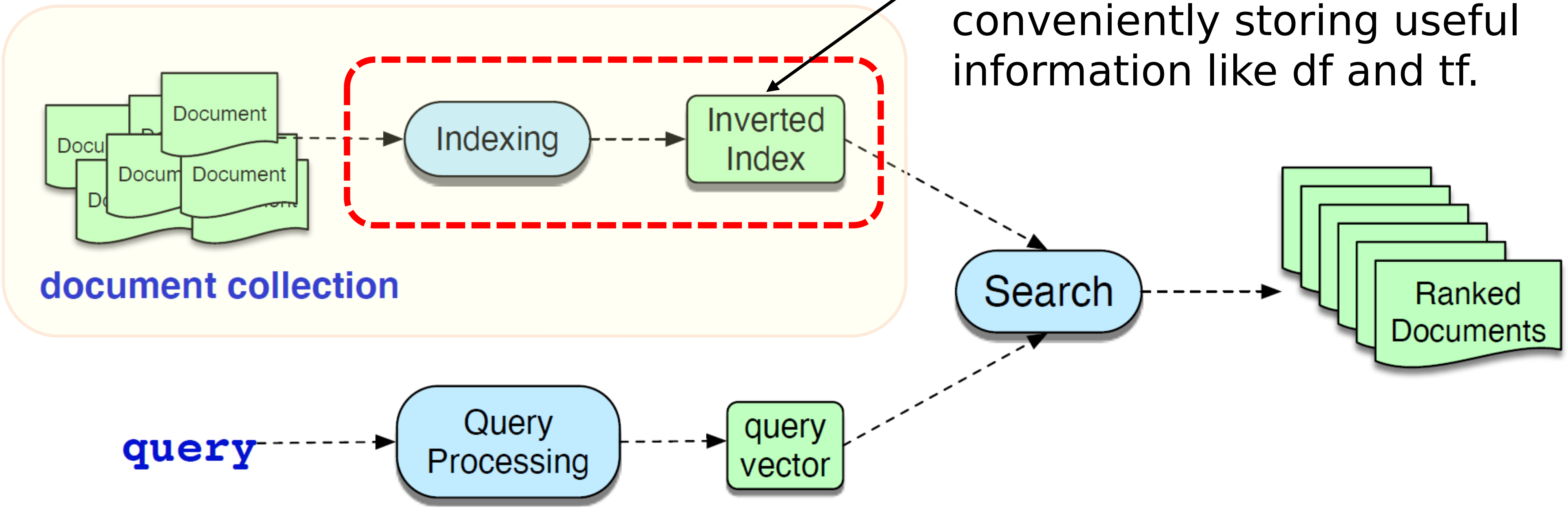
[3] Document Rank

Ranked Retrieval Result	*d1*	*d4*	*d2*	

Text Ranking

- **Information Retrieval**

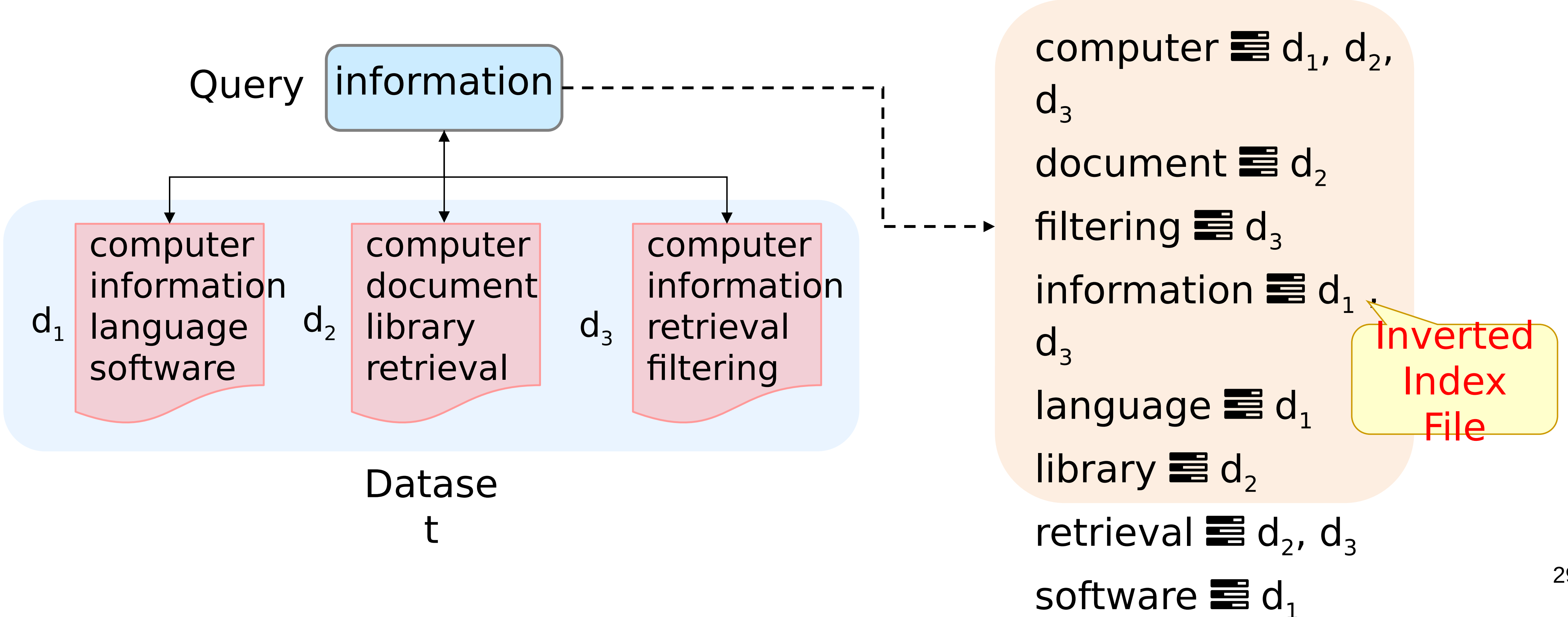
It is the data structure used for making search efficient and also conveniently storing useful information like df and tf.



Architecture of Information Retrieval

Text Ranking

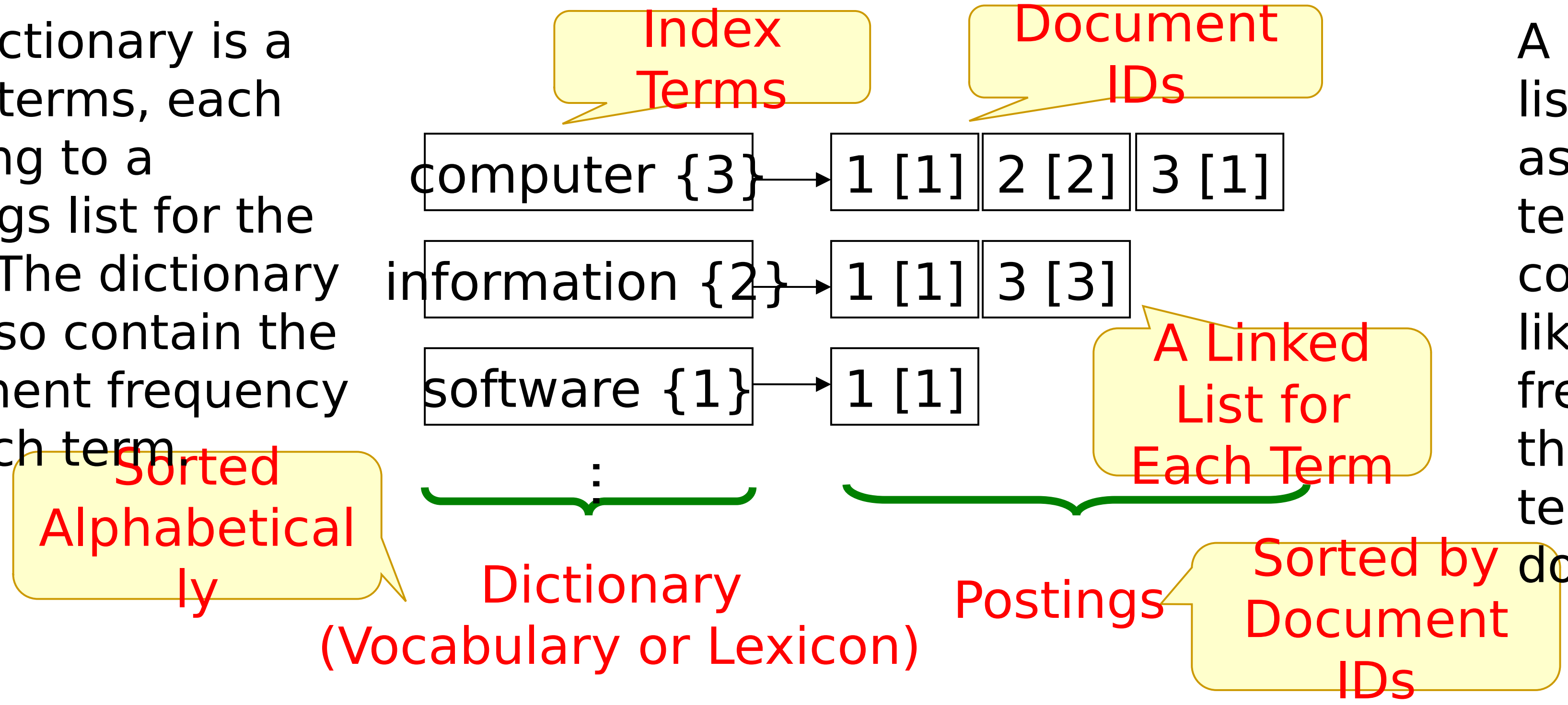
- **Information Retrieval**
  - An inverted index is a list of documents that contain the term.



Text Ranking

- **Information Retrieval**
  - The inverted index consists of two parts, a dictionary and the postings.

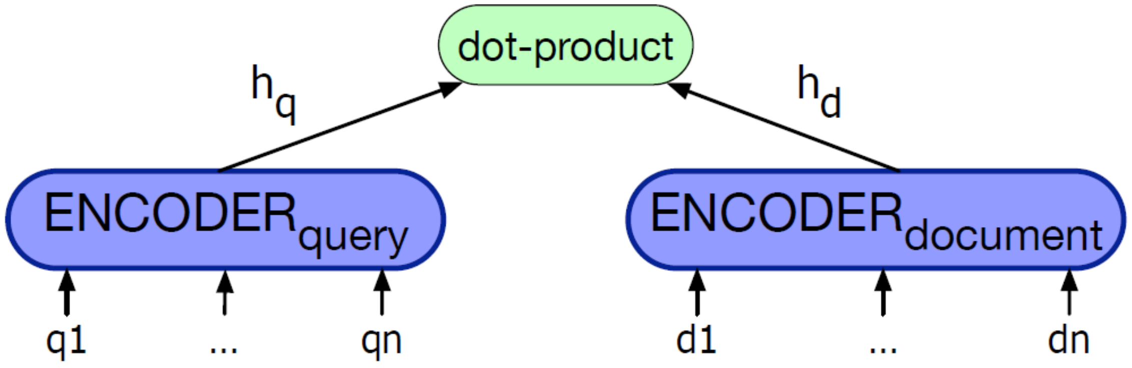
The dictionary is a list of terms, each pointing to a postings list for the term. The dictionary can also contain the document frequency for each term.



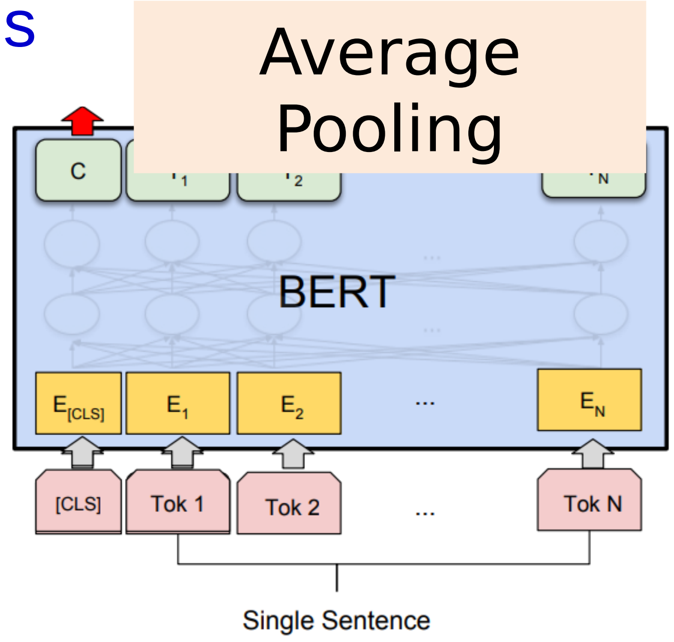
A postings list is the list of document IDs associated with each term, which can also contain information like the term frequency or even the exact positions of terms in the document.

Text Ranking

- **Dense Retrieval**
  - Vocabulary Mismatch Problem: Synonyms
  - From Sparse BOW Vectors to Dense Embeddings



BERT Bi-Encoder for Computing Revelance of a Document to a Query

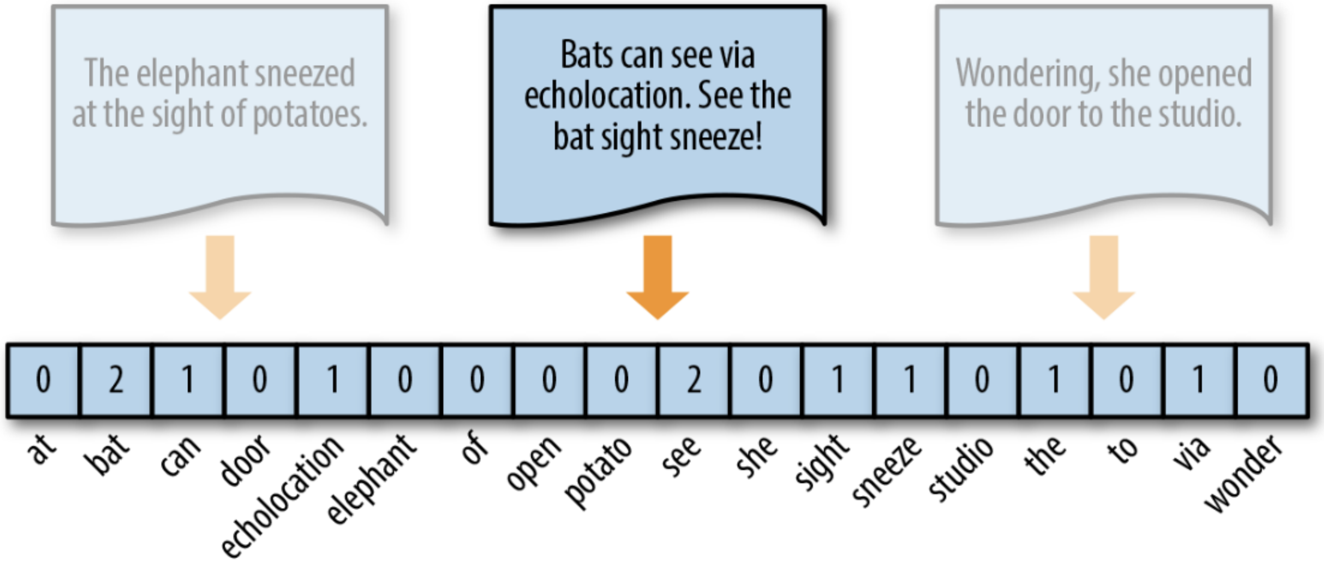


- BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding

Text Ranking

- **Dense Retrieval**
  - From Sparse BOW Vectors to Dense Embeddings

BOW Representation



Embedding

0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1	0.7	0.4	0.3

References

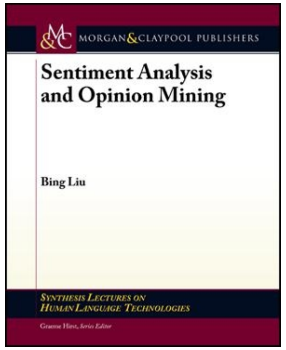
- **Book Chapters**
  - Speech and Language Processing
    - Chapter 4: Naive Bayes, Text Classification and Sentiment (NB)
    - Chapter 5: Logistic Regression (LR)
    - Chapter 14.1: Information Retrieval

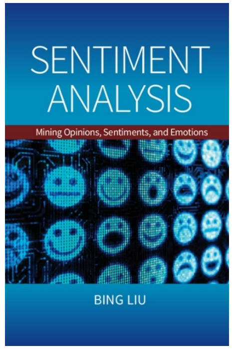
References

- **Book Chapters**
  - Introduction to Information Retrieval
    - Chapter 6: Scoring, Term Weighting and the Vector Space Model
    - Chapter 11: Probabilistic Information Retrieval
    - Chapter 12: Language Models for Information Retrieval
    - Chapter 13: Text Classification and Naive Bayes (NB)
    - Chapter 14: Vector Space Classification (*k*-NN)
    - Chapter 15: Support Vector Machines and Machine Learning on Documents (SVM)

References

- **Reference Books**
  - [Sentiment Analysis and Opinion Mining]
  - [Sentiment Analysis: Mining Sentiments, Opinions, and Emotions]





Announcement

- **Lab 1**
  - Venue: PQ604A/B/C
  - Time: 6:30pm ~ 9:20pm
  - Date: Tuesday, February 11, 2025
  - Tutor: Heming Xia


thank you