Master Degree in Computer Science
Master Degree in Data Science and Economics

Information Retrieval

Scoring, term weighting and the vector space model

Prof. Alfio Ferrara

Department of Computer Science, Università degli Studi di Milano
Room 7012 via Celoria 18, 20133 Milano, Italia alfio.ferrara@unimi.it

sed noli modo

Text transformation

In order to process queries and other IR tasks on text, we need to transform it into a model that can be computed by a machine

Such a model should be

- suitable for any kind of text
- capable of modeling interesting properties of text, such as the semantics
- suitable for matching texts and supporting the notion of *text similarity* in a quantitative framework

The vector space model (intuition)

Doc 1

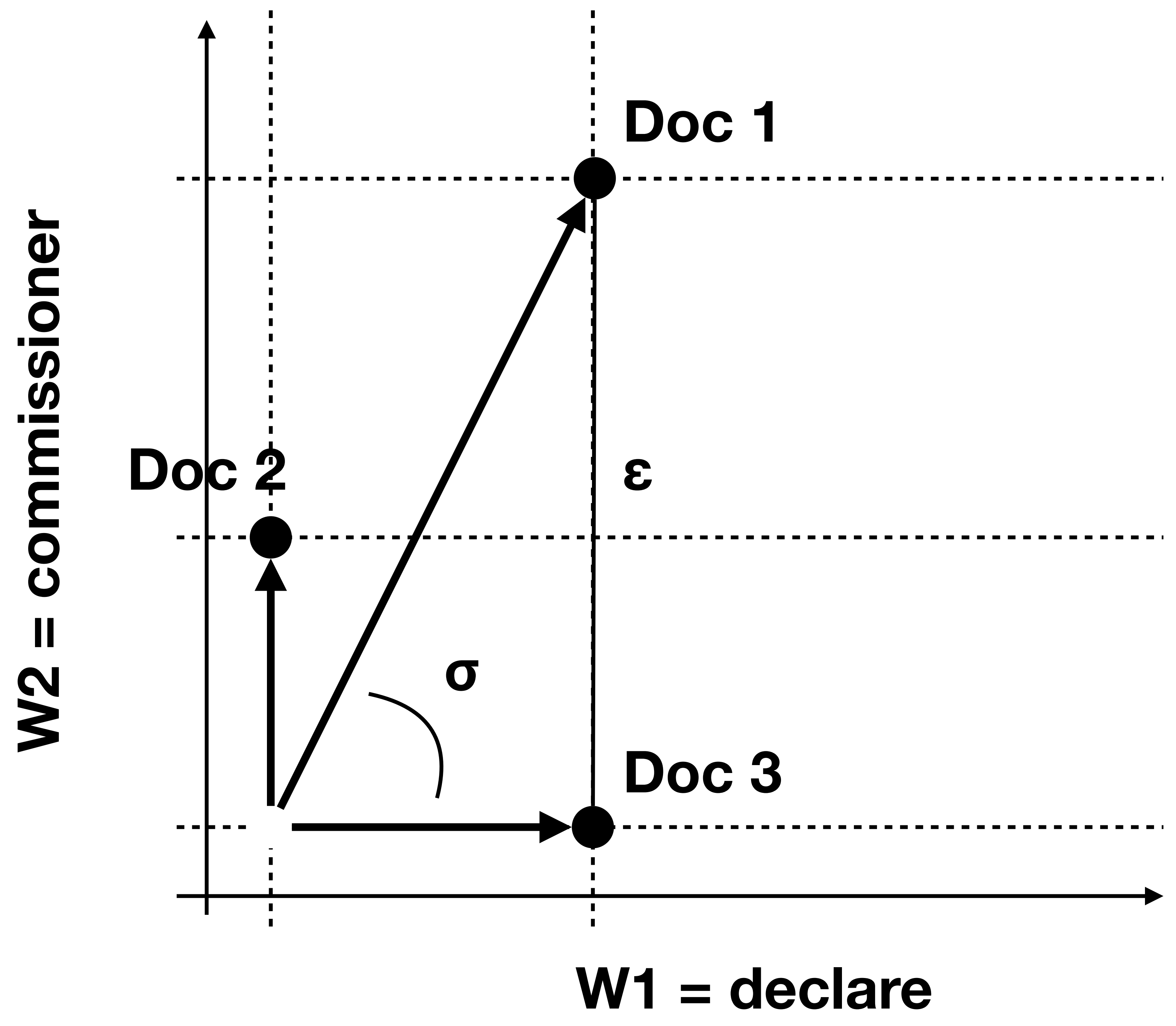
TO REVISE THE CHARTER; Governor Soon to Announce His Choice of Commissioners. [...] The Commissioners declared that [...]

Doc 2

JAMES McCARTNEY DEAD.; The Commissioner of Street Cleaning Passes Away at His Home After a Long Illness.

Doc 3

DENY COLGAN LOST JOB OVER 'AL' SMITH; Hylan Declares Cry of Politics Shows Attempt to Split Democratic Party.



Doc	W1	W2	...
Doc1	1	2	...
Doc 2	0	1	...
Doc 3	1	0	...

Mapping documents into the vector space

Doc 1

TO REVISE THE CHARTER; Governor Soon to Announce His Choice of Commissioners. [...] The Commissioners declared that [...]

TOKENIZE

```
['TO', 'REVISE', 'THE', 'CHARTER', ';', 'Governor', 'Soon', 'to', 'Announce', 'His', 'Choice',
'of', 'Commissioners', '.', 'The', 'Commissioners', 'declared', 'that']
```

NORMALIZE

```
['to', 'revis', 'the', 'charter', ';', 'governor', 'soon', 'to', 'announc', 'hi', 'choic', 'of',
'commission', '.', 'the', 'commission', 'declar', 'that']
```

WEIGHT

```
{'to': 2, 'the': 2, 'commission': 2, 'revis': 1, 'charter': 1, ';': 1, 'governor': 1, 'soon': 1,
'announc': 1, 'hi': 1, 'choic': 1, 'of': 1, '.': 1, 'declar': 1, 'that': 1}
```

INDEXING

	to	the	commission	revis	charter	;	governor	soon	announc	hi	choic	of	.	declar	that
Doc1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1

Tokenize

REGEX	https://www.nltk.org/_modules/nltk/tokenize/regexp.html
CORPUS-BASED	https://www.nltk.org/api/nltk.tokenize.html
ML MODELS & RULE-BASED	https://spacy.io/usage/linguistic-features#tokenization

Normalize

STEMMING

"The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems."

M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137.

DICTIONARY-BASED LEMMATIZATION

WordNet: https://wordnet.princeton.edu

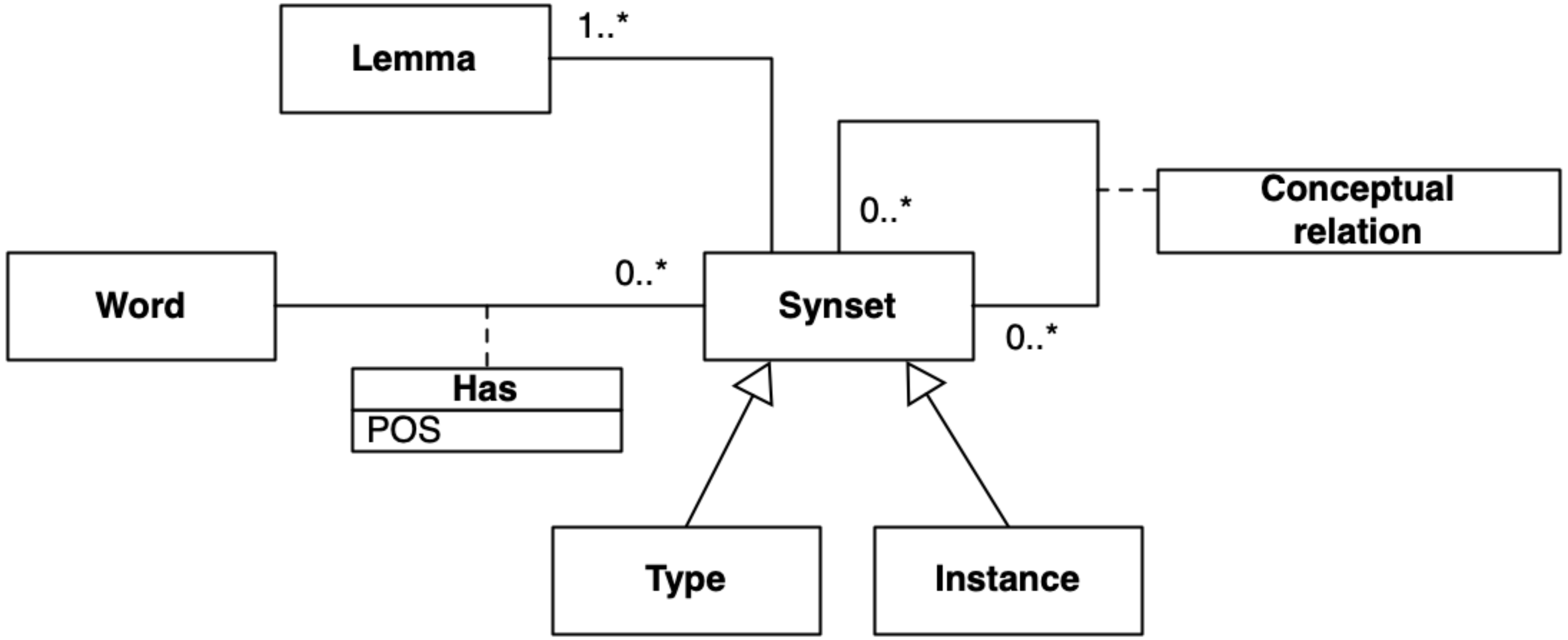
ML-MODELS

https://spacy.io/usage/linguistic-features#lemmatization

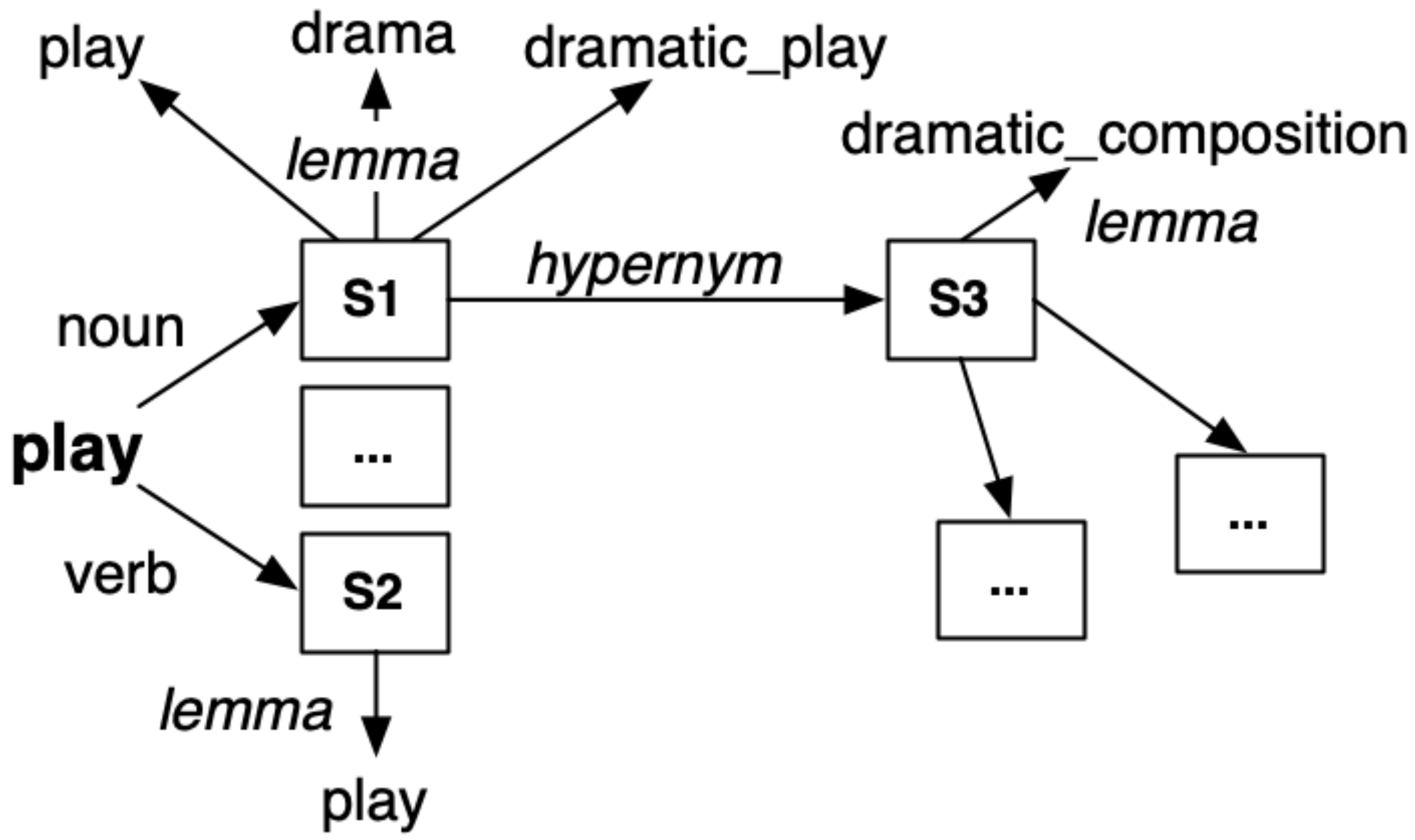
WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

MODEL



EXAMPLE



Weight: Term Frequency (TF)

BOOLEAN $1 \ if \ tf_{t,d} > 0, \ 0 \ otherwise$

AUGMENTED $0.5 + \frac{0.5tf_{t,d}}{max_{t'}(tf_{t'd})}$

NATURAL $tf_{t,d} = count(t) \in d$

LOG AVG $\frac{1 + \log(tf_{t,d})}{1 + \log(avg_{t \in d}(tf_{t,d}))}$

LOG $1 + \log(tf_{t,d})$

MAX TF NORM $k + (1 - k)\frac{tf_{t,d}}{tf_{max}(d)}$

Weight: Inverse Document Frequency (IDF)

Besides term frequency, another important measure for weighting terms is the **document frequency**. The document frequency $df(t)$ of a term t is given by the number of documents d_i where $tf_{t,d_i} > 0$.

In many applications, we are interested in terms that are specific of a small subset of documents and, thus, have a low document frequency. To this end, we define the inverse document frequency idf_t as

$$idf_t = \log \frac{N}{df_t}$$

Weight: Inverse Document Frequency (IDF)

IDF $$\log \frac{N}{df_t} = -\log \frac{df_t}{N}$$

MAX $$\log \left(\frac{\max_{t' \in d} df_{t'}}{1 + df_t} \right) + 1$$

SMOOTH $$\log \left(\frac{N}{1 + df_t} \right) + 1$$

PROBABILISTIC $$\log \frac{N - df_t}{df_t}$$

Weight: TfIdf

By combining term frequency and inverse document frequency we obtain a measure that takes into account the specific relevance of a terms with respect to a document

$$TfIdf(t, d) = tf_{t,d} idf_t = tf_{t,d} \log \frac{N}{df_t}$$

TfIdf is:
- high when t appears in a small number of documents (high discriminating)
- low when t appears a few times in *d* or when it appears in many documents (irrelevant or generic)

Exploit vectors to measure document distances

Distance	Definition	Interpretation	Distance	Definition	Interpretation
Cosine Distance	$\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_1^2}\sqrt{\sum_{i=1}^{n} b_1^2}}$	Cosine of the angle between the vectors	Canberra distance	$\sum_{i=1}^{n} \frac{\mid a_i - b_i \mid}{\mid a_i \mid \mid b_i \mid}$	Normalized version of Manhattan distance
Euclidean Distance	$\sqrt{\sum_{i=1}^{n} \mid a_i - b_i \mid^2}$	Vector points distance	Bray-Curtis distance	$\frac{\sum_{i=1}^{n} \mid a_i - b_i \mid}{\sum_{i=1}^{n} \mid a_i + b_i \mid}$	Variant of Manhattan distance
Chebyshev distance	$\max_{i=1}^{n}(\mid a_i - b_i \mid)$	The greatest different along any dimension	Correlation distance	$\frac{\sum_{i=1}^{n} (a_i - avg\ a)(b_i - avg\ b)}{\sqrt{\left\lvert \sum_{i=1}^{n} (a_i - avg\ a) \right\rvert^2} \sqrt{\left\lvert \sum_{i=1}^{n} (b_i - avg\ b) \right\rvert^2}}$	Equivalent to Cosine Distance of vectors shifted by their means
Manhattan distance	$\sum_{i=1}^{n} \mid a_i - b_i \mid$	Distance in a grid	Minkowski distance	$\left(\sum_{i=1}^{n} \mid a_i - b_i \mid^p \right)^{\frac{1}{p}}$	Generalization of Manhattan (p=1) and Euclidean (p=2) distances