Data Mining
(Mining Knowledge from Data)

Text and Web Mining

Marcel Jiřina, Pavel Kordík

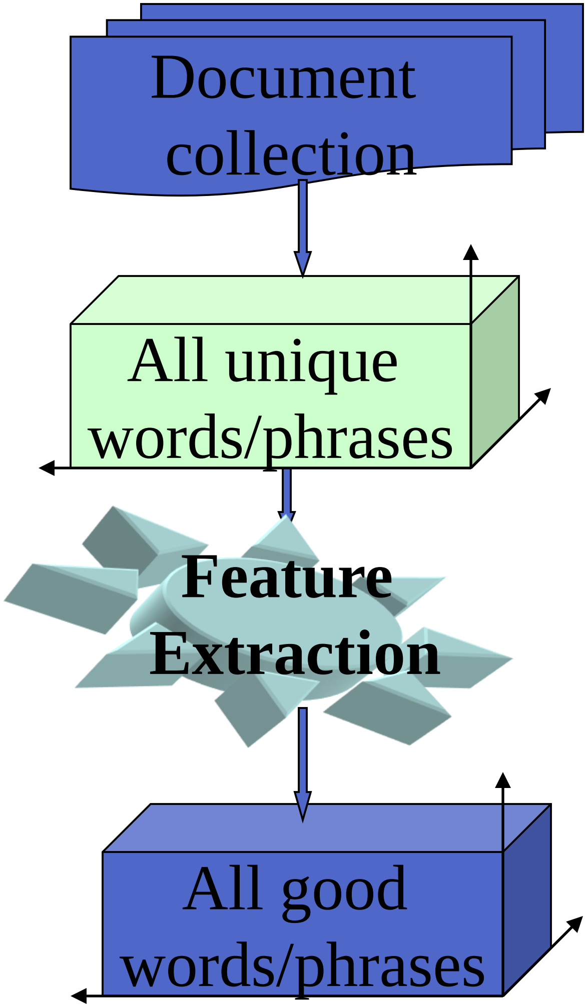
ČESKÉ
VYSOKÉ
UČENÍ
TECHNICKÉ
V PRAZE

FIT

Feature Extraction from text

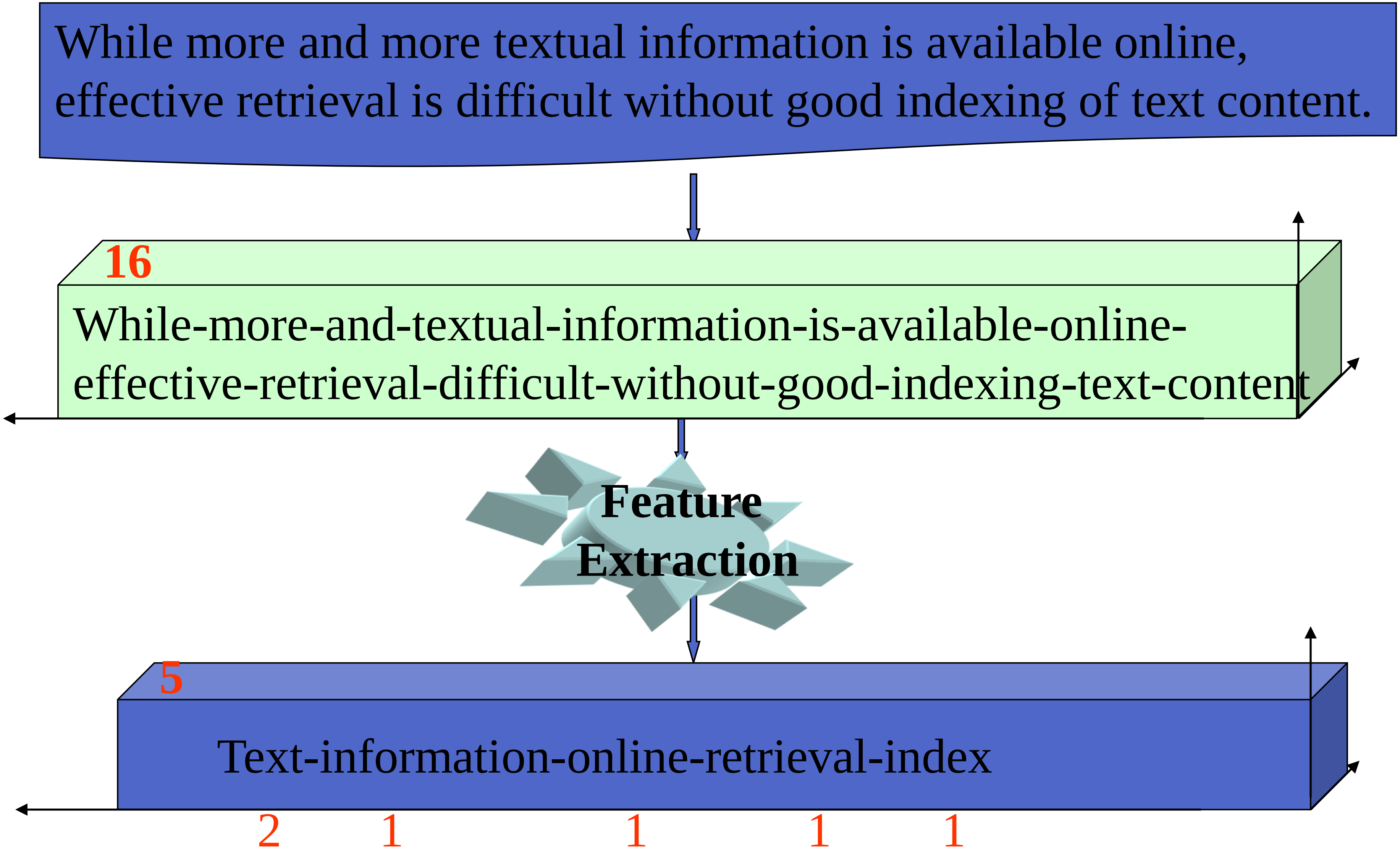
- Basic
  - Representative words
  - Indexing
  - Weighting Model
  - Dimensionality Reduction
- Linguistic
  - Part-of-speech tagging
  - Syntactic parsing

Feature Extraction: Representative words



Task: Extract a good subset of words to represent documents

Representative words extraction, example



While more and more textual information is available online, effective retrieval is difficult without good indexing of text content.

16

While-more-and-textual-information-is-available-online-effective-retrieval-difficult-without-good-indexing-text-content

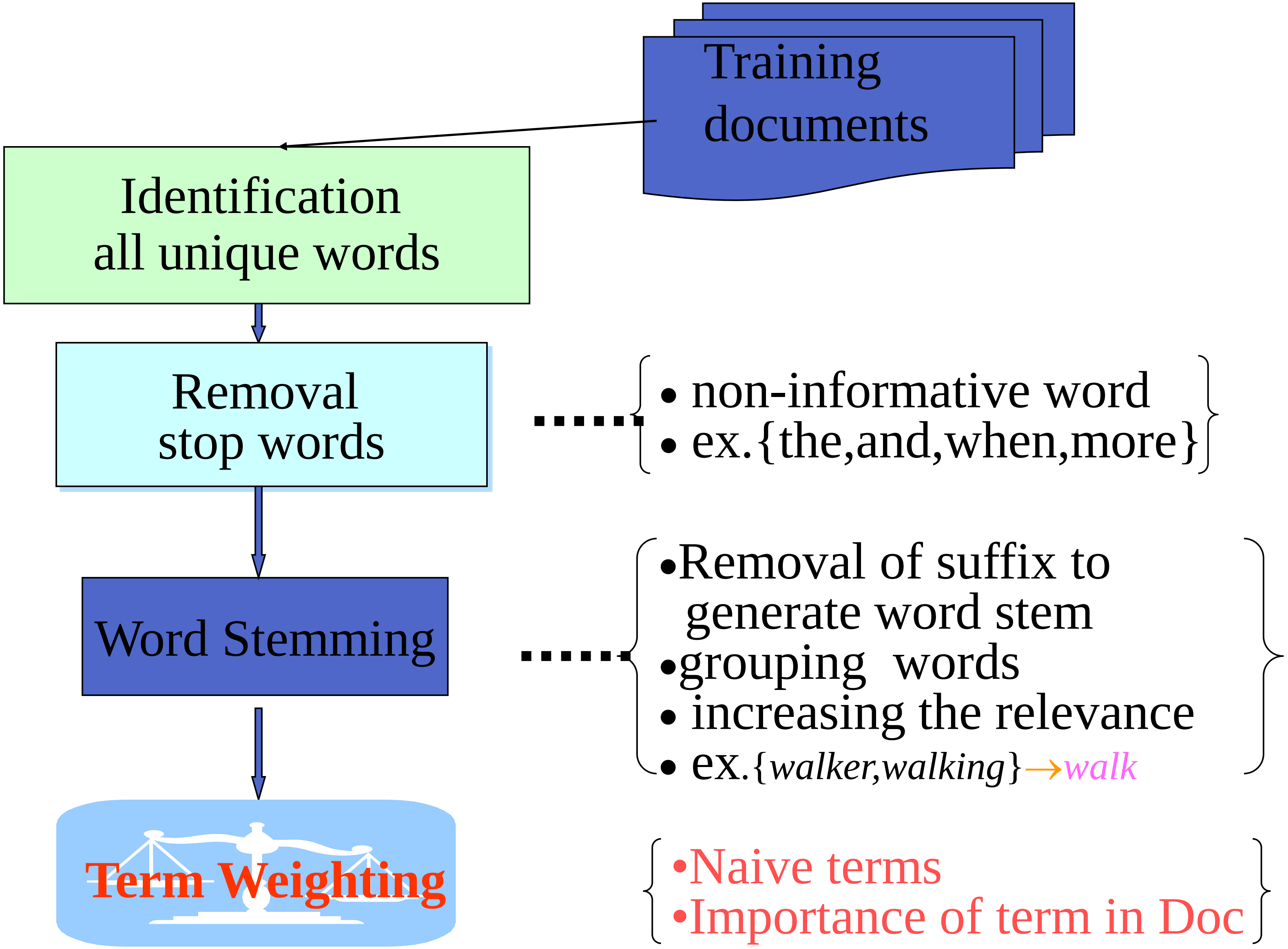
Feature Extraction

5

Text-information-online-retrieval-index

2 1 1 1 1

Feature Extraction: Indexing



Training documents

Identification all unique words

Removal stop words

- non-informative word
- ex.{the,and,when,more}

Word Stemming

- Removal of suffix to generate word stem
- grouping words
- increasing the relevance
- ex.{*walker,walking*}→*walk*

Term Weighting

- Naive terms
- Importance of term in Doc

Feature Extraction: Indexing(2)

- Document representations: vector space models

$$d = (w_1, w_2, \ldots w_t) \in \mathbf{R}^t$$

w_i is the weight of *i*th term in document *d*.

Word: sleep

↓

{sle, lee, eep}

- set of 3grams

- Terms can be words or **ngrams**
  - prefer ngrams because of the need to process noisy and/or multilingual documents

Feature Extraction: Weighting Model(tf)

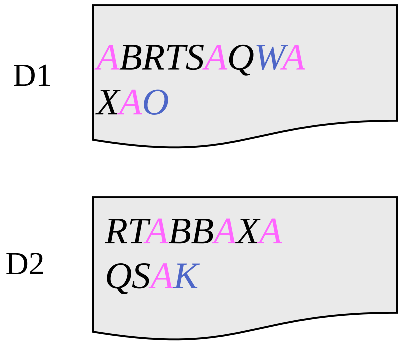
- **tf - Term Frequency weighting**

  w_{ij} = Freq$_{ij}$

  Freq$_{ij}$ = the number of times jth term occurs in document D_i.

  × Drawback: without reflection of importance factor for document discrimination.

- **Ex.**



D1

ABRTSAQWA
XAO

D2

RTABBAXA
QSAK

	A	B	K	O	Q	R	S	T	W	X
D1	3	1	0	1	1	1	1	1	1	1
D2	3	2	1	0	1	1	1	1	0	1

Feature Extraction:Weighting Model(**tf×idf**)

- **tf×idf - Inverse Document Frequency weighting**

  $\mathbf{w_{ij}} = \mathbf{Freq_{ij}} * \log(\mathbf{N} / \mathbf{DocFreq_j})$.

  N = the number of documents in the training document collection.

  $\mathbf{DocFreq_j}$ = the number of documents in which the jth term occurs.

  ✓Advantage: with reflection of importance factor for document discrimination.

  Assumption:terms with low DocFreq are better discriminator than ones with high DocFreq in document collection

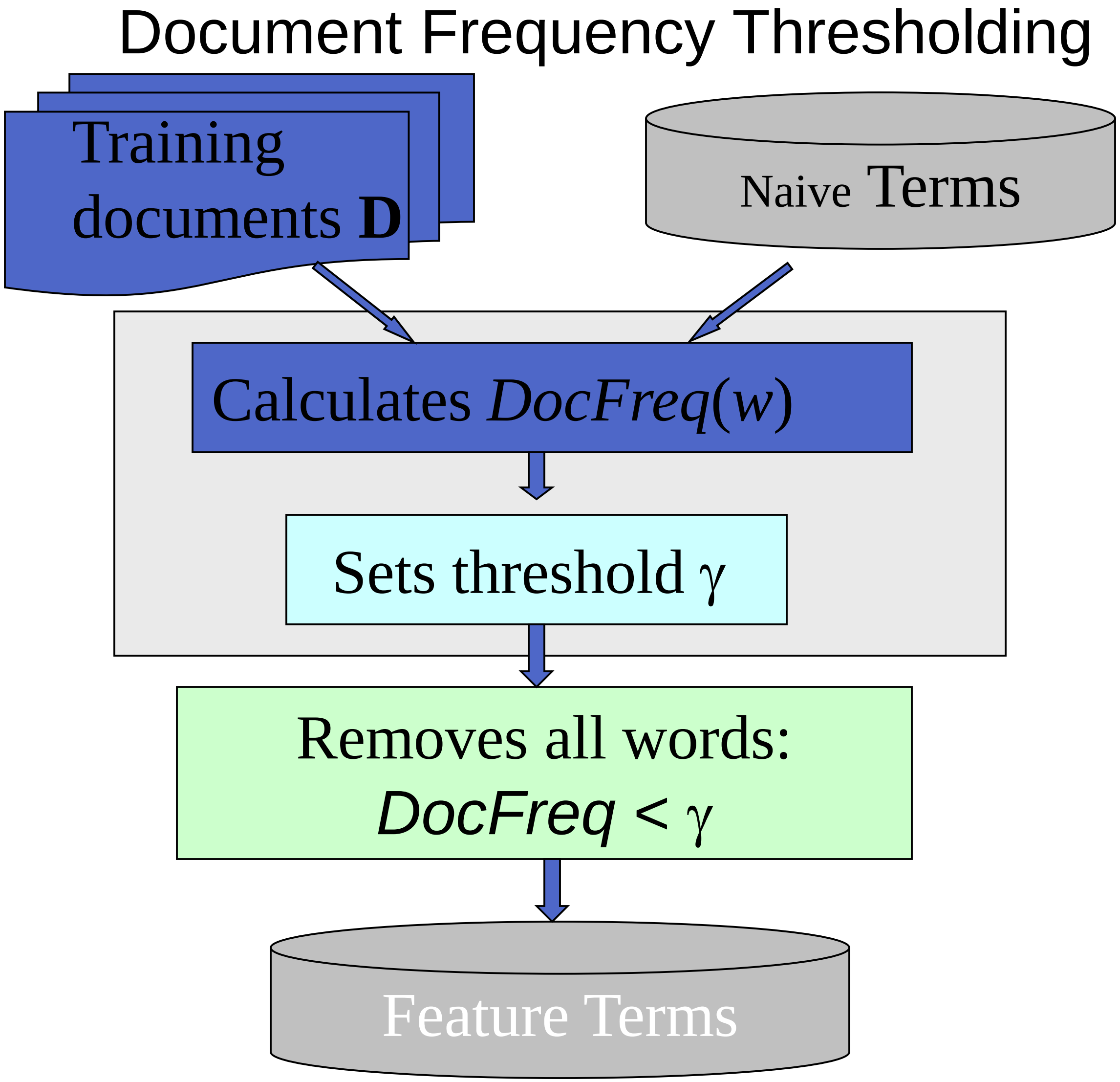
- Ex.

	A	B	K	O	Q	R	S	T	W	X
D1	0	0	0	0.3	0	0	0	0	0.3	0
D2	0	0	0.3	0	0	0	0	0	0	0

Feature Extraction: Dimension Reduction

- Document Frequency Thresholding
- X^2-statistic
- Latent Semantic Indexing

Dimension Reduction: DocFreq Thresholding


Document Frequency Thresholding

Training documents **D**

Naive Terms

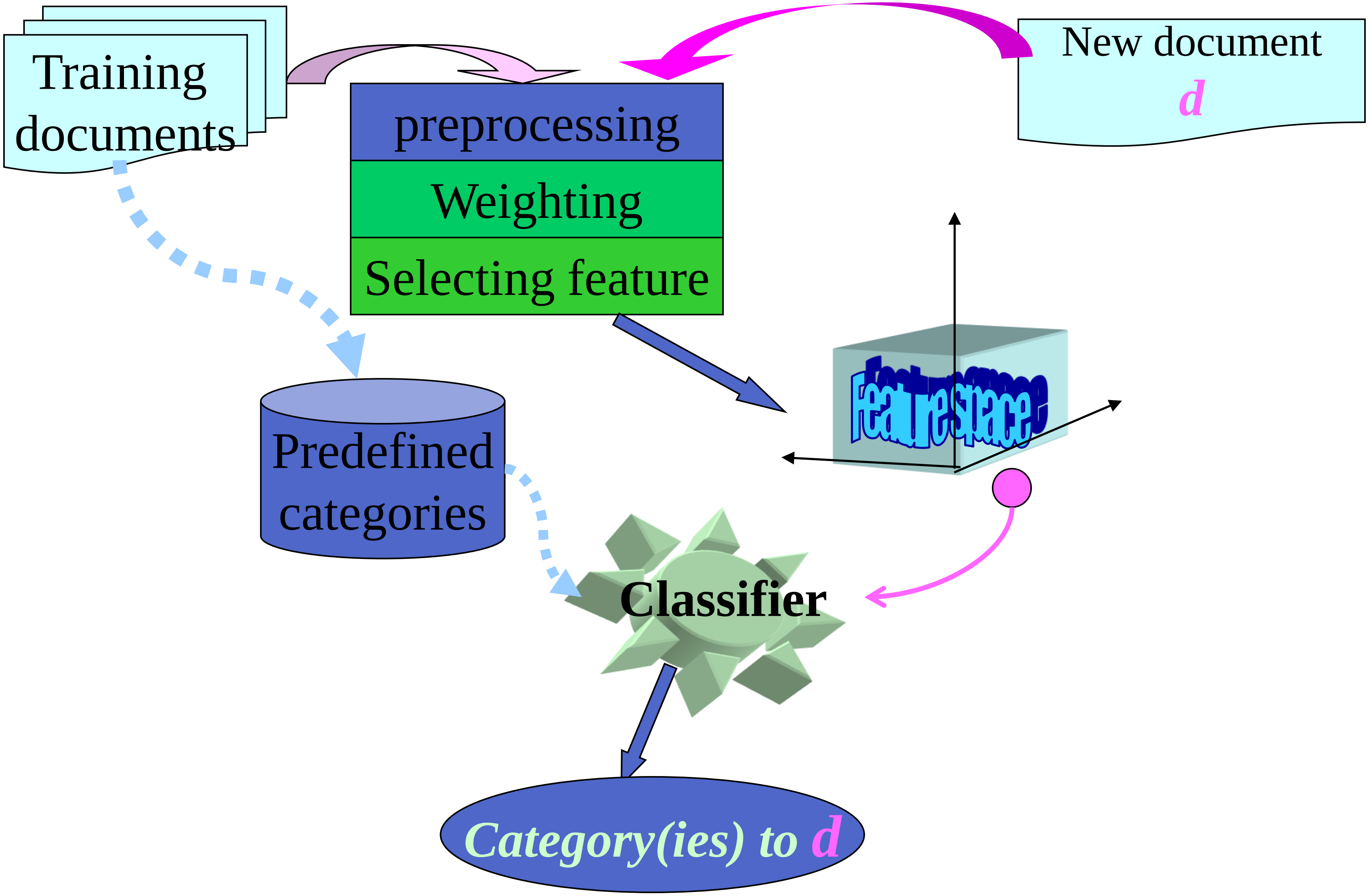
Calculates *DocFreq*(*w*)

Sets threshold γ

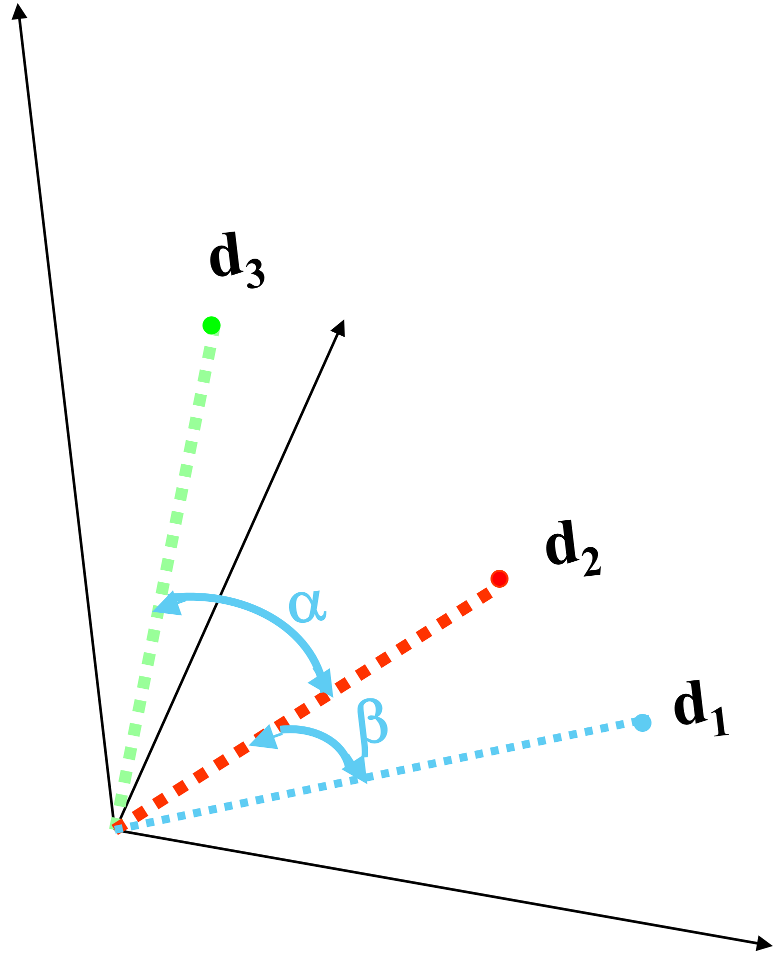
Removes all words: *DocFreq* < γ

Feature Terms

Document Categorization: Architecture



3.3.1 Model:Centroid-Based Classifier(2)



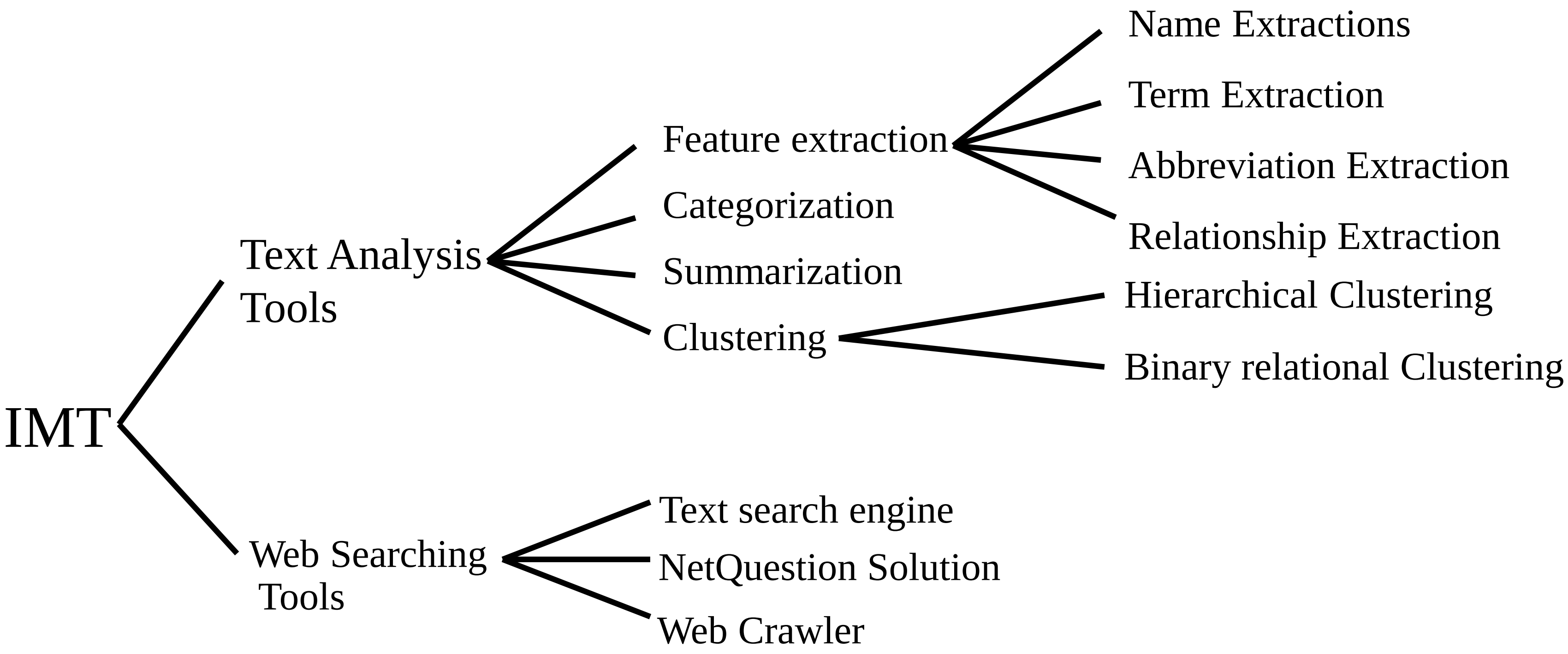
- α > β
- cos(α)<cos(β)
- **d_2** is more close to **d_1** than **d_3**

$d = (w_1, w_2, \ldots, w_n);$

$$\cos(d_i, d_j) = \frac{d_i \bullet d_j}{\|d_i\|_2 \times \|d_j\|_2}$$

Cosine-based similarity model can reflect the *relations between features*.

Intelligent Miner for Text(IMT)(1)



IMT

Text Analysis Tools

Feature extraction

Name Extractions

Term Extraction

Abbreviation Extraction

Relationship Extraction

Categorization

Summarization

Clustering

Hierarchical Clustering

Binary relational Clustering

Web Searching Tools

Text search engine

NetQuestion Solution

Web Crawler

Intelligent Miner for Text(IMT)(2)

Feature extraction tools

1.1 Information extraction

- Extract linguistic items that represent document contents

1.2 Feature extraction

- Assign of different categories to vocabulary in documents,
- Measure their importance to the document content.

1.3 Name extraction

- Locate names in text,
- Determine what type of entity the name refers to

1.4 Term extraction

- Discover terms in text. Multiword technical terms
- Recognize variants of the same concept

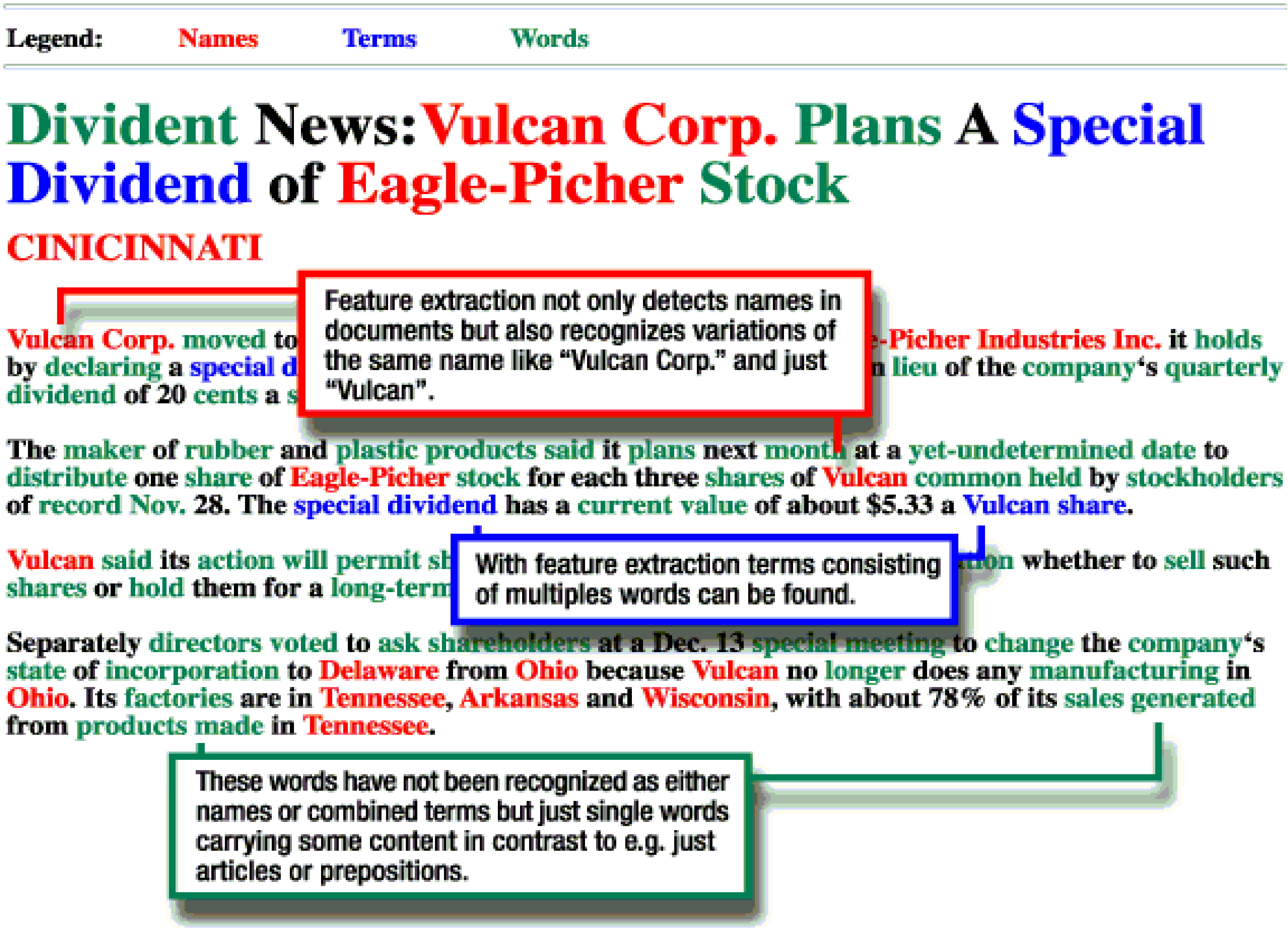
1.5 Abbreviation recognition

- Find abbreviation and math them with their full forms.

1.6 Relation extraction

Intelligent Miner for Text(IMT)(3)

Feature extraction Demo



Legend: Names Terms Words

Divident News:Vulcan Corp. Plans A Special Dividend of Eagle-Picher Stock

CINICINNATI

Feature extraction not only detects names in documents but also recognizes variations of the same name like “Vulcan Corp.” and just “Vulcan”.

Vulcan Corp. moved to -Picher Industries Inc. it holds by declaring a special d n lieu of the company‘s quarterly dividend of 20 cents a s

The maker of rubber and plastic products said it plans next month at a yet-undetermined date to distribute one share of Eagle-Picher stock for each three shares of Vulcan common held by stockholders of record Nov. 28. The special dividend has a current value of about $5.33 a Vulcan share.

With feature extraction terms consisting of multiples words can be found.

Vulcan said its action will permit sh ion whether to sell such shares or hold them for a long-term

Separately directors voted to ask shareholders at a Dec. 13 special meeting to change the company‘s state of incorporation to Delaware from Ohio because Vulcan no longer does any manufacturing in Ohio. Its factories are in Tennessee, Arkansas and Wisconsin, with about 78% of its sales generated from products made in Tennessee.

These words have not been recognized as either names or combined terms but just single words carrying some content in contrast to e.g. just articles or prepositions.

Intelligent Miner for Text(IMT)(4)

Clustering tools

Applications

- Provide a overview of content in a large document collection
- Identify hidden structures between groups of objects
- Improve the browsing process to find similar or related information
- Find outstanding documents within a collection

Hierarchical clustering

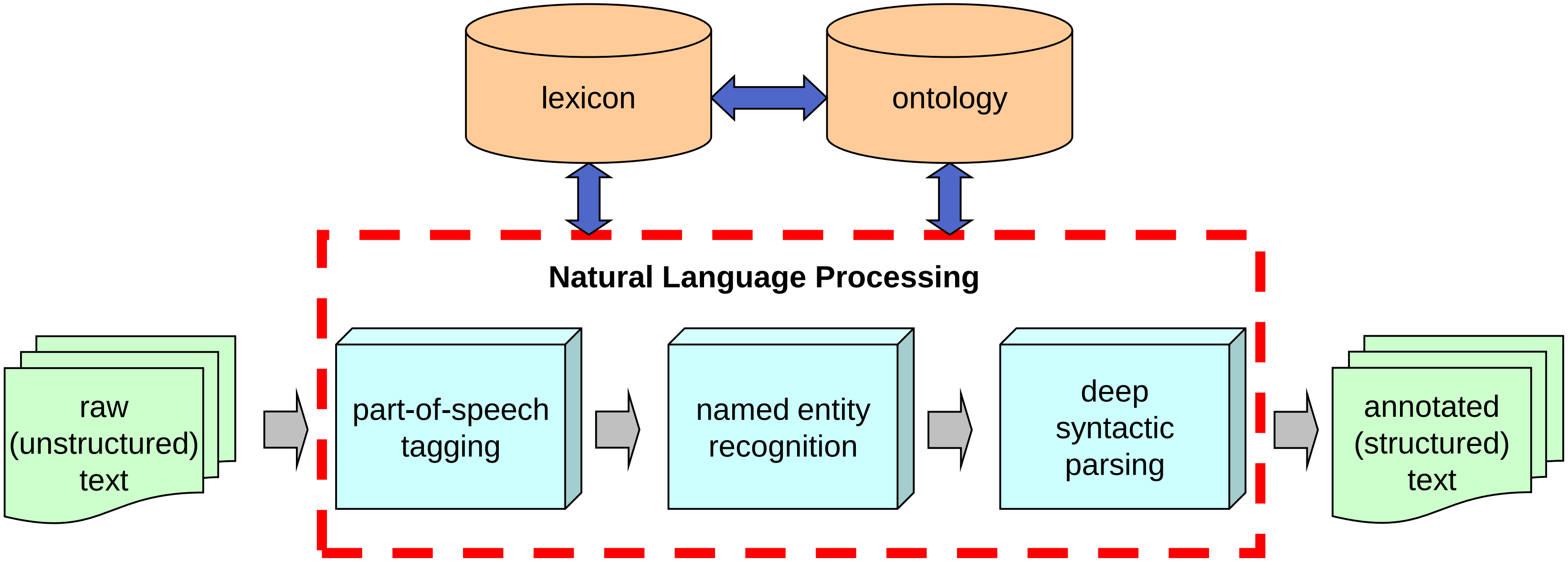
- Clusters are organized in a clustering tree and related clusters occurs in the same branch of tree.

Binary relational clustering

- Relationship of topics.
- document → cluster →topic.

Linguistic features extraction

- Known as Natural Language Processing (NLP)
- Different for English, Czech …



Basic Steps of Natural Language Processing

- Sentence splitting
- Tokenization
- Part-of-speech tagging
- Shallow parsing
- Named entity recognition
- Syntactic parsing
- (Semantic Role Labeling)

Sentence splitting

Current immunosuppression protocols to prevent lung transplant rejection reduce pro-inflammatory and T-helper type 1 (Th1) cytokines. However, Th1 T-cell pro-inflammatory cytokine production is important in host defense against bacterial infection in the lungs. Excessive immunosuppression of Th1 T-cell pro-inflammatory cytokines leaves patients susceptible to infection.

Current immunosuppression protocols to prevent lung transplant rejection reduce pro-inflammatory and T-helper type 1 (Th1) cytokines.

However, Th1 T-cell pro-inflammatory cytokine production is important in host defense against bacterial infection in the lungs.

Excessive immunosuppression of Th1 T-cell pro-inflammatory cytokines leaves patients susceptible to infection.

A heuristic rule for sentence splitting

sentence boundary

= period + space(s) + capital letter

Regular expression in Perl

```
s/\. +([A-Z])/\.\n\1/g;
```

Errors

IL-33 is known to induce the production of Th2-associated cytokines (e.g. IL-5 and IL-13).

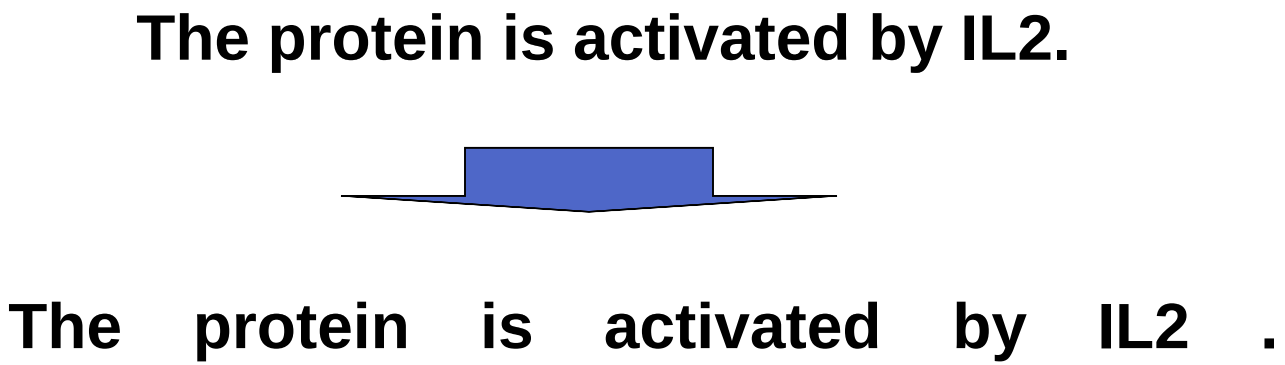
↓

IL-33 is known to induce the production of Th2-associated cytokines (e.g.

IL-5 and IL-13).

- Two solutions:
  - – Add more rules to handle exceptions
  - – Machine learning

Tokenization



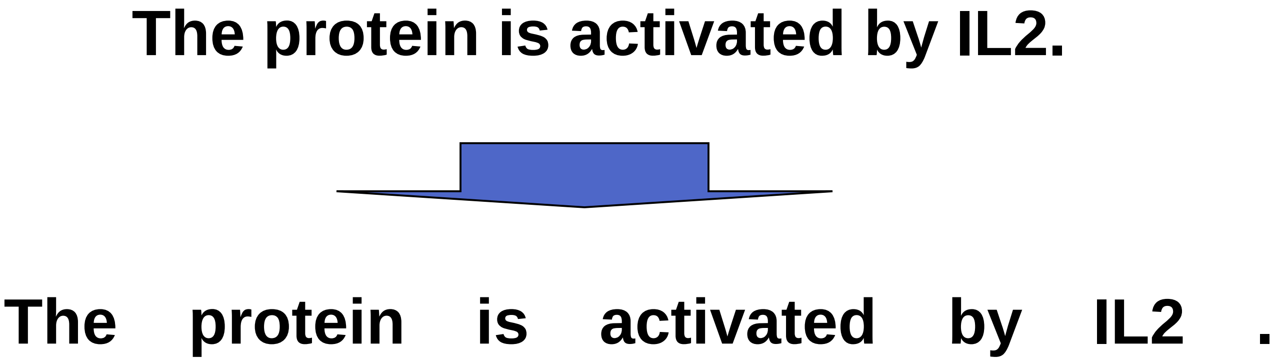
The protein is activated by IL2.

The protein is activated by IL2 .

- Convert a sentence into a sequence of *tokens*

- Why do we tokenize?
- Because we do not want to treat a sentence as a sequence of *characters*!

Tokenization



The protein is activated by IL2.

The protein is activated by IL2 .

- Tokenizing general English sentences is relatively straightforward.
- Use spaces as the boundaries
- Use some heuristics to handle exceptions

Tokenization issues

- separate possessive endings or abbreviated forms from preceding words:
  - Mary’s → Mary ‘s
    Mary’s → Mary is
    Mary’s → Mary has
- separate punctuation marks and quotes from words :
  - Mary. → Mary .
  - “new” → “ new “

Tokenization

- Tokenizer.sed: a simple script in *sed*
  - http://www.cis.upenn.edu/~treebank/tokenization.html
- **Tokenization:** Divides the text into smallest units (usually words), removing punctuation.
- Challenge: What should be done with punctuation that has linguistic meaning?

Part-of-speech tagging

The peri-kappa B site mediates human immunodeficiency
DT NN NN NN VBZ JJ NN
virus type 2 enhancer activation in monocytes …
NN NN CD NN NN IN NNS

- Assign a part-of-speech tag to each token in a sentence.

Part-of-speech tags

- The Penn Treebank tagset
  - http://www.cis.upenn.edu/~treebank/
  - 45 tags

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
:	:
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBZ	Verb, 3rd person singular present
:	:

JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
:	:
DT	Determiner
CD	Cardinal number
CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
FW	Foreign word
:	:

Part-of-speech tagging is not easy

- Parts-of-speech are often ambiguous

I have to go to school.

verb

I had a go at skiing.

noun

- We need to look at the context
- But how?

Writing rules for part-of-speech tagging

I have to go to school.
verb

I had a go at skiing.
noun

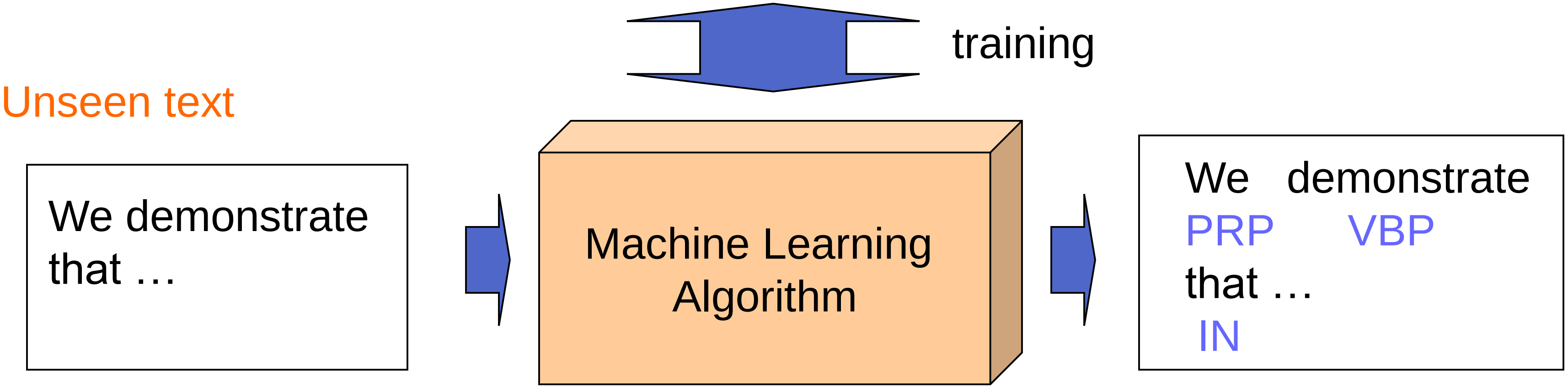
- If the previous word is “to”, then it’s a verb.
- If the previous word is “a”, then it’s a noun.
- If the next word is …

:

⇒ **Writing rules manually is impossible**

Learning from examples

The involvement of ion channels in B and T lymphocyte activation is
DT NN IN NN NNS IN NN CC NN NN NN VBZ
supported by many reports of changes in ion fluxes and membrane
VBN IN JJ NNS IN NNS IN NN NNS CC NN
……………………………………………………………………………………
……………………………………………………………………………………



training

Unseen text

We demonstrate that …

Machine Learning Algorithm

We demonstrate
PRP VBP
that …
IN

Tagging errors made by a WSJ-trained POS tagger

… and membrane potential after mitogen ~~binding~~.
CC NN NN IN NN JJ

… two factors, which ~~bind~~ to the same kappa B enhancers…
CD NNS WDT NN TO DT JJ NN NN NNS

… by analysing the ~~Ag~~ amino acid sequence.
IN VBG DT VBG JJ NN NN

… to contain ~~more~~ T-~~cell~~ determinants than …
TO VB RBR JJ NNS IN

Stimulation of interferon beta gene transcription ~~in vitro~~ by
NN IN JJ JJ NN NN IN NN IN