

# Text and Web Mining

Marija Stanojevic

Knowledge Discovery and Data Mining Course  
Temple University

22<sup>nd</sup> March 2018

# Text Mining

- Finding non-trivial, hidden, unknown and useful patterns in large textual datasets
- Intersection of many areas
- Easy:
  - There are simple and good algorithms for simple tasks
  - Highly redundant data
- Hard:
  - Abstract concepts difficult to represent / visualize
  - High dimensionality



# Levels of Processing

- Character Level
- Word Level
- Sentence Level
- Document Level
- Document – Collection Level
- Linked – Document – Collection Level

# Character Level Processing

- Much slower than word level processing for natural language
- Much less accurate than word level processing for natural language
- Used in DNA / RNA analysis
- Similar analysis are used as in word level processing, but longer sequences are required to achieve accuracy for natural language

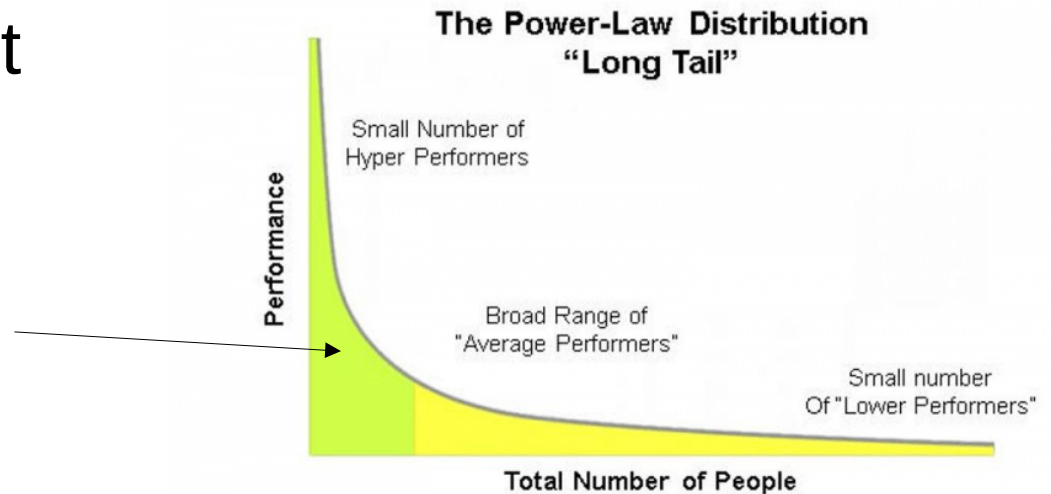
# Word Level Processing (1)

- **Words properties**

- **Homonymy**: same form, different meaning (bank: river bank, financial institution)
- **Polysemy**: same form, related meaning (bank: blood bank, financial institution)
- **Synonymy**: different form, same meaning (singer, vocalist)
- **Hyponymy**: word denotes subclass of other (breakfast, meal)
- Frequencies in text have power law distribution

- **Stop-words removal**

- Language dependent
- Removes green part



# Word Level Processing (2)

- **Stemming**

- Different forms of the same word are problematic for many algorithms
- Heuristics and rules for transforming a word into a normalized form (ex. cats → cat, eating, ate → eat)
- Many stemmers (Porter stemmer mostly used)

- **Tokenization**

- Splitting by words (can be split by characters)

- **N-grams**

- Sequence of n consecutive tokens

- **WordNet**

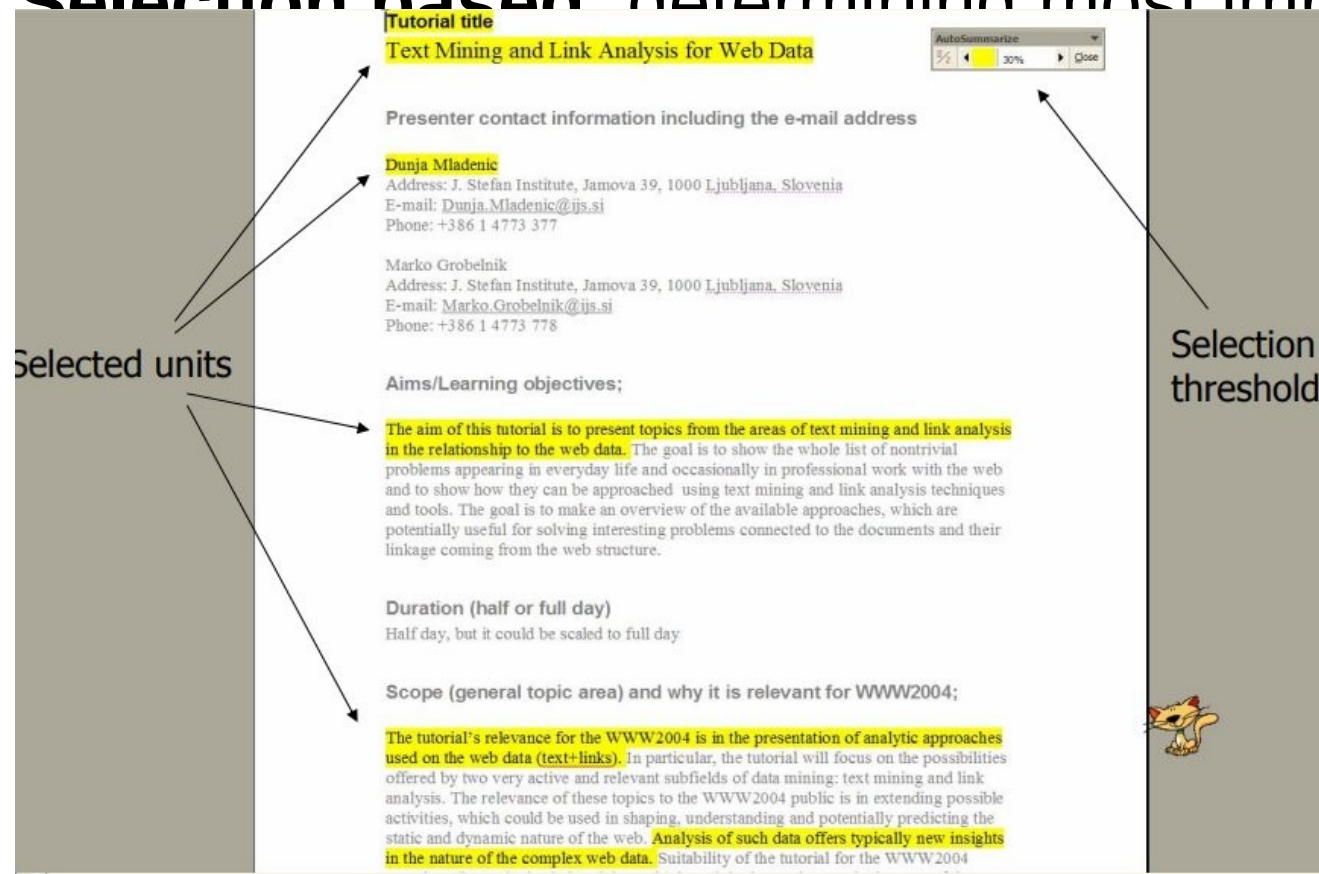
- Lexical database for English

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

# Document Level Processing (1)

- **Summarization**

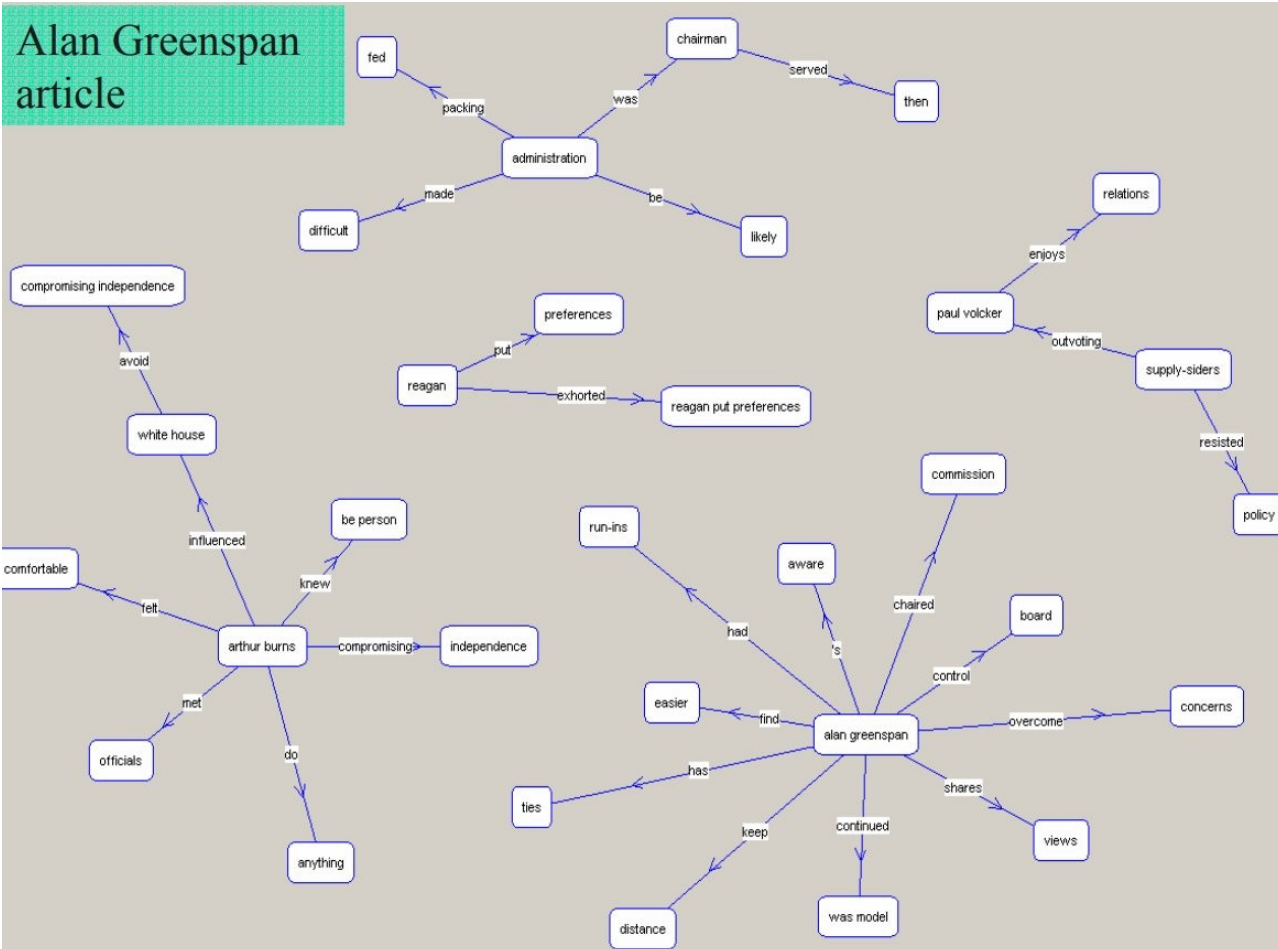
- **Knowledge rich:** performing semantic analysis, representing the meaning and generating the text satisfying length restriction
- **Selection based:** determining most important parts



# Document Level Processing (2)

- **Visualization**

- can't count on statistical properties (especially for short documents)
- Use syntactical and logical structure





# Document-Collection Level Processing (3)

## The Bag of Words Representation

- **Representation**

- Bag of words
- Word weighting

- Tf-idf
- Glove
- Latent Semantic Analysis (LSA)

- Statistical representation

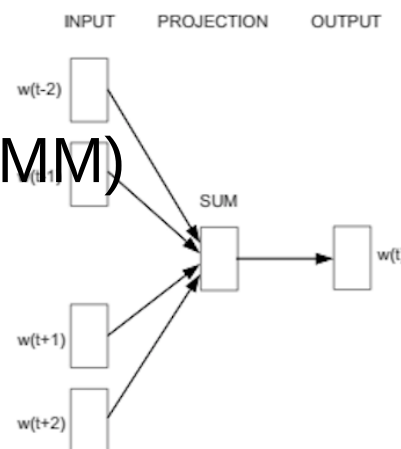
- Latent Dirichlet Analysis (LDA)
- Dirichlet Multinomial Mixture (DMM)
- Word-Topic Model (WTM)

- Word vectors
- word2vec

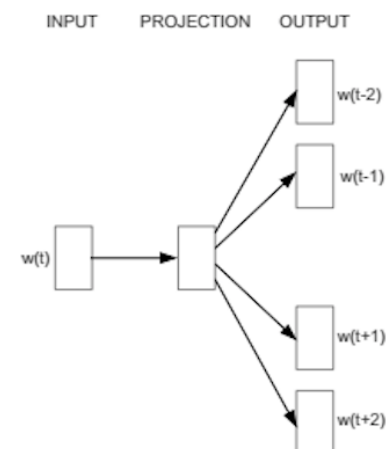
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy 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
...	...



CBOW



Skip-gram

# Document-Collection Level Processing (4)

- **Classification**

- Supervised learning of document labels / topics
- General classifiers of document vectors
- RNN/LSTM classification of documents based on their text sequences
- Statistical classification (LDA, WTM)

- **Clustering**

- Unsupervised learning of document topics
- General clustering of document vectors

# Document-Collection Level Processing (5)

- **Visualization**

- Cluster documents by topics and visualize clusters
- Transform word vectors in 2D and visualize
- Transform documents vectors in 2D and visualize

- **Information Extraction**

- Named entity extraction
- Classification of entities
- Association extraction
- Knowledge database creation

October 14, 2002, 4:00 a.m. PT

For years, **Microsoft Corporation** **CEO Bill Gates** railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, **Microsoft** claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. **Gates** himself says **Microsoft** will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

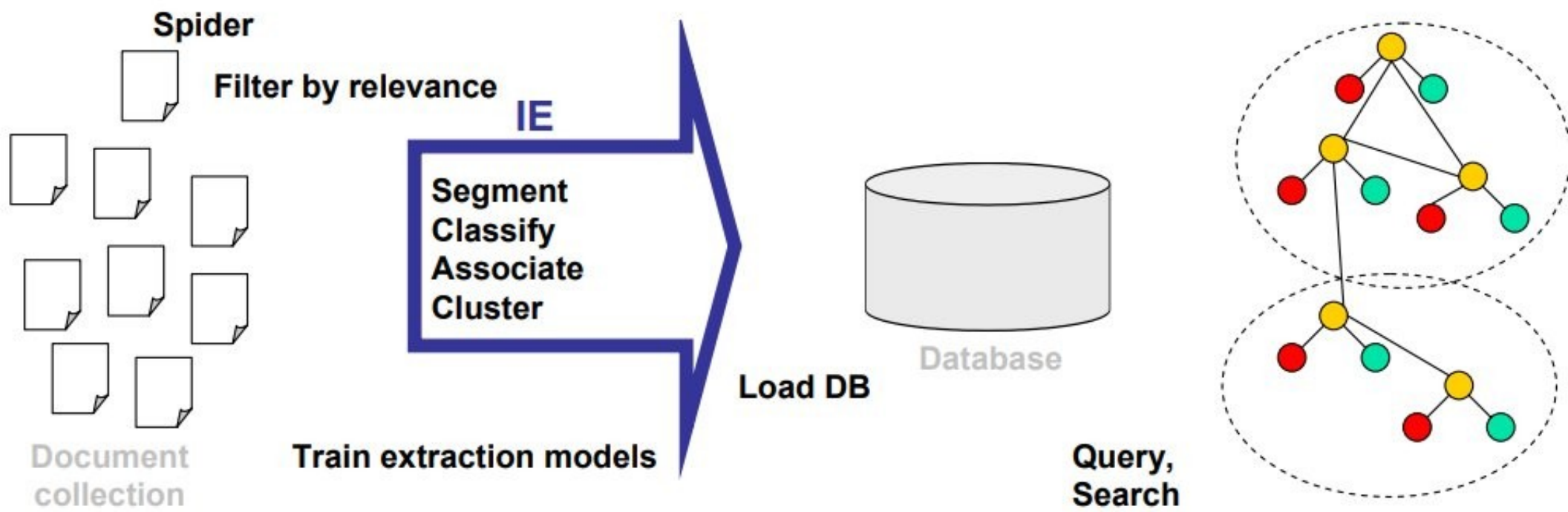
"We can be open source. We love the concept of shared source," said **Bill Veghte**, a **Microsoft VP**. "That's a super-important shift for us in terms of code access."

**Richard Stallman**, **founder** of the **Free Software Foundation**, countered saying...

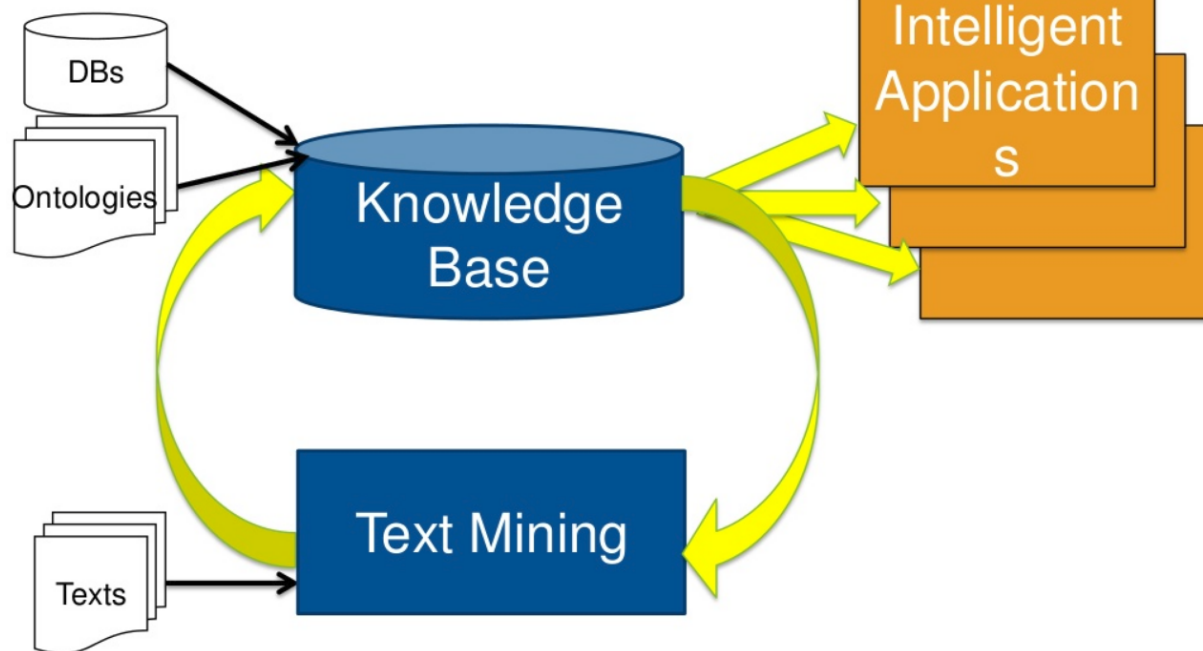
*	<b>Microsoft Corporation</b>
	<b>CEO</b>
	<b>Bill Gates</b>
*	<b>Microsoft</b>
	<b>Gates</b>
*	<b>Microsoft</b>
	<b>Bill Veghte</b>
*	<b>Microsoft</b>
	<b>VP</b>
	<b>Richard Stallman</b>
	<b>founder</b>
	<b>Free Software Foundation</b>

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

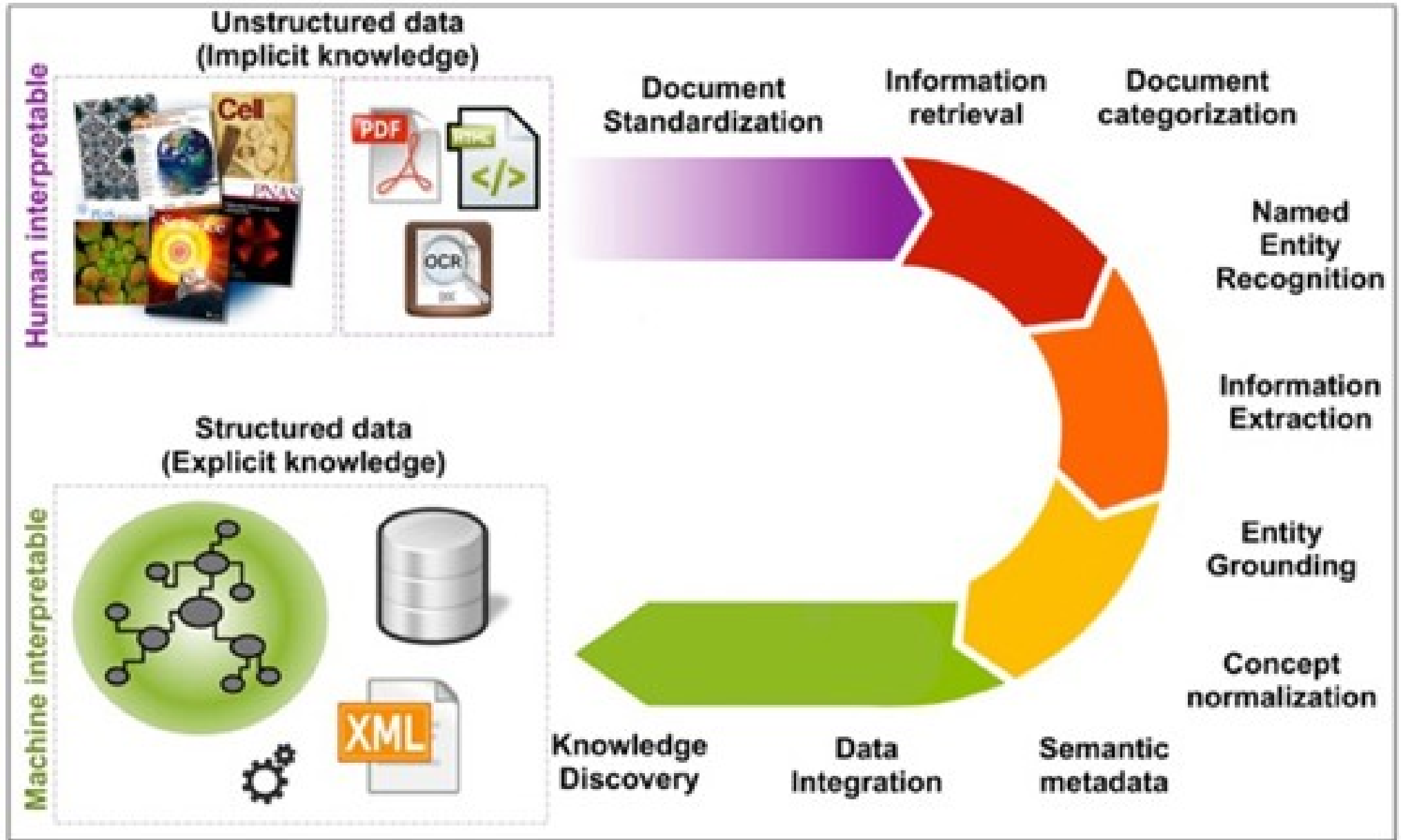
## Create ontology



## Label training data



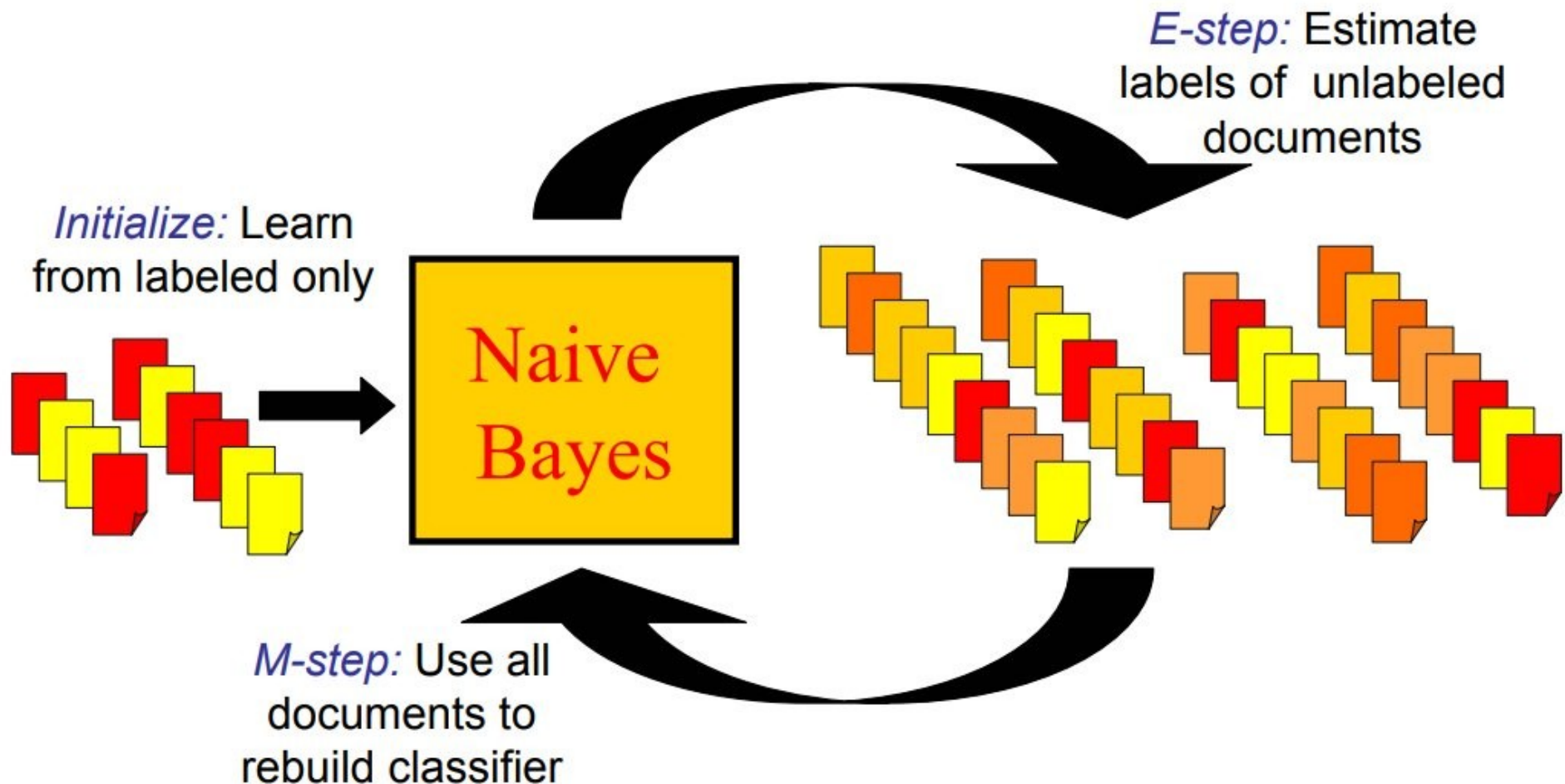
# Example of knowledge extraction





# Linked-Document-Collection Level (1)

## Using Unlabeled Data with Expectation-Maximization (EM)



Guarantees local maximum a posteriori parameters

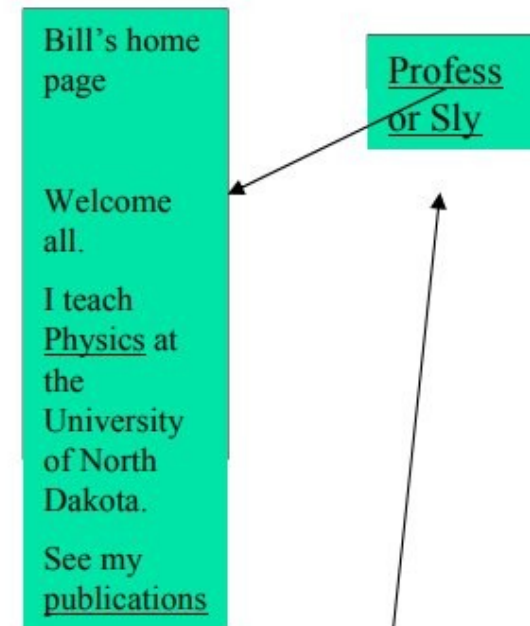
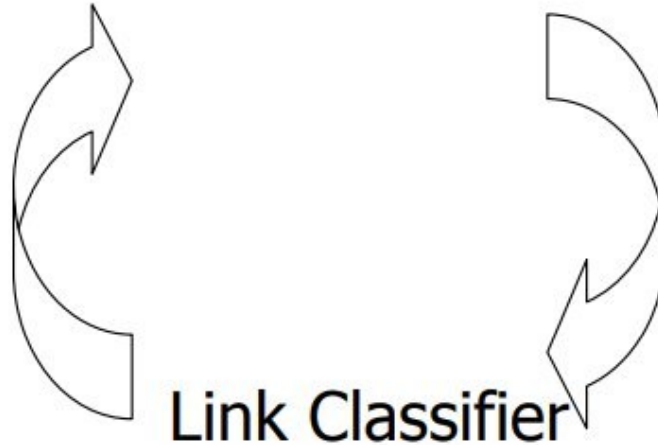
# Linked-Document-Collection Level (2)

## Bootstrap Learning to Classify Web Pages (co-training)

**Given:** set of documents where each document is described by two independent sets of attributes (e.g. text + hyperlinks)

12 labeled pages

Page Classifier



Hyperlink, pointing to the document

Document content

# References

## 1)Text Mining Tutorial:

[http://www.dainf.ct.utfpr.edu.br/~kaestner/Mineracao/RDataMining/Tutorial\\_Marko.pdf](http://www.dainf.ct.utfpr.edu.br/~kaestner/Mineracao/RDataMining/Tutorial_Marko.pdf)

2)Witten, I. H. (2004). Text Mining.

3)Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.

4)Jiang, J., & Allan, J. (2017, November). Similarity-based Distant Supervision for Definition Retrieval. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 527-536). ACM.

5)Wijaya, D., Talukdar, P. P., & Mitchell, T. (2013, October). Pidgin: ontology alignment using web text as interlingua. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 589-598). ACM.

6)Wijaya, D. T., Nakashole, N., & Mitchell, T. (2015). " A Spousal Relation Begins with a deletion of engage and Ends with an Addition of divorce": Learning State Changing Verbs from Wikipedia Revision History. In Proceedings of the 2015 conference on empirical methods in natural language processing(pp. 518-523).

7)Wijaya, D. T., & Mitchell, T. M. (2016). Mapping verbs in different languages to knowledge base relations using web text as interlingua. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 818-827).

8)Wijaya, D. T., Callahan, B., Hewitt, J., Gao, J., Ling, X., Apidianaki, M., & Callison-Burch, C. (2017). Learning Translations via Matrix Completion. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1452-1463).



# Homework

**Document 1:** Norway is building a city that is totally powered by renewable energy.

**Document 2:** Three countries will power two-thirds of global renewable energy growth.

**Document 3:** New York will invest \$1.5 billion in renewable energy projects.

## Task:

- Change all text to lower-case, remove punctuation, remove stop-words and do stemming
- Tokenize and create tf-idf for given three documents
- Calculate cosine similarity between each two documents based on tf-idf
- Does the sentences have more similar meaning in your opinion than the score says?

## Example:

- **Document:** These two countries are the reason because of which the EU is hitting its ambitious renewable energy targets.
- **Lower case:** these two countries are the reason because of which the eu is hitting its ambitious renewable energy targets
- **Remove stop-words:** two countries are reason eu is hitting its ambitious renewable energy targets
- **After stemming:** two country be reason eu be hit its ambitious renewable energy target
- **After tokenization and tf:** two (1), country (1), be (2), reason (1), hit (1), its (1), ambitious (1), renewable energy (1) target (1)
- **Further details:** <https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/>