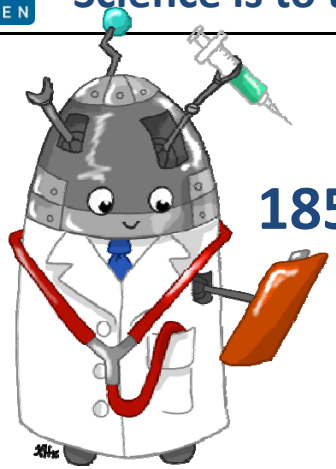


Andreas Holzinger

185.A83 Machine Learning for Health Informatics

2016S, VU, 2.0 h, 3.0 ECTS

Week 25 - 22.06.2016 17:00-20:00

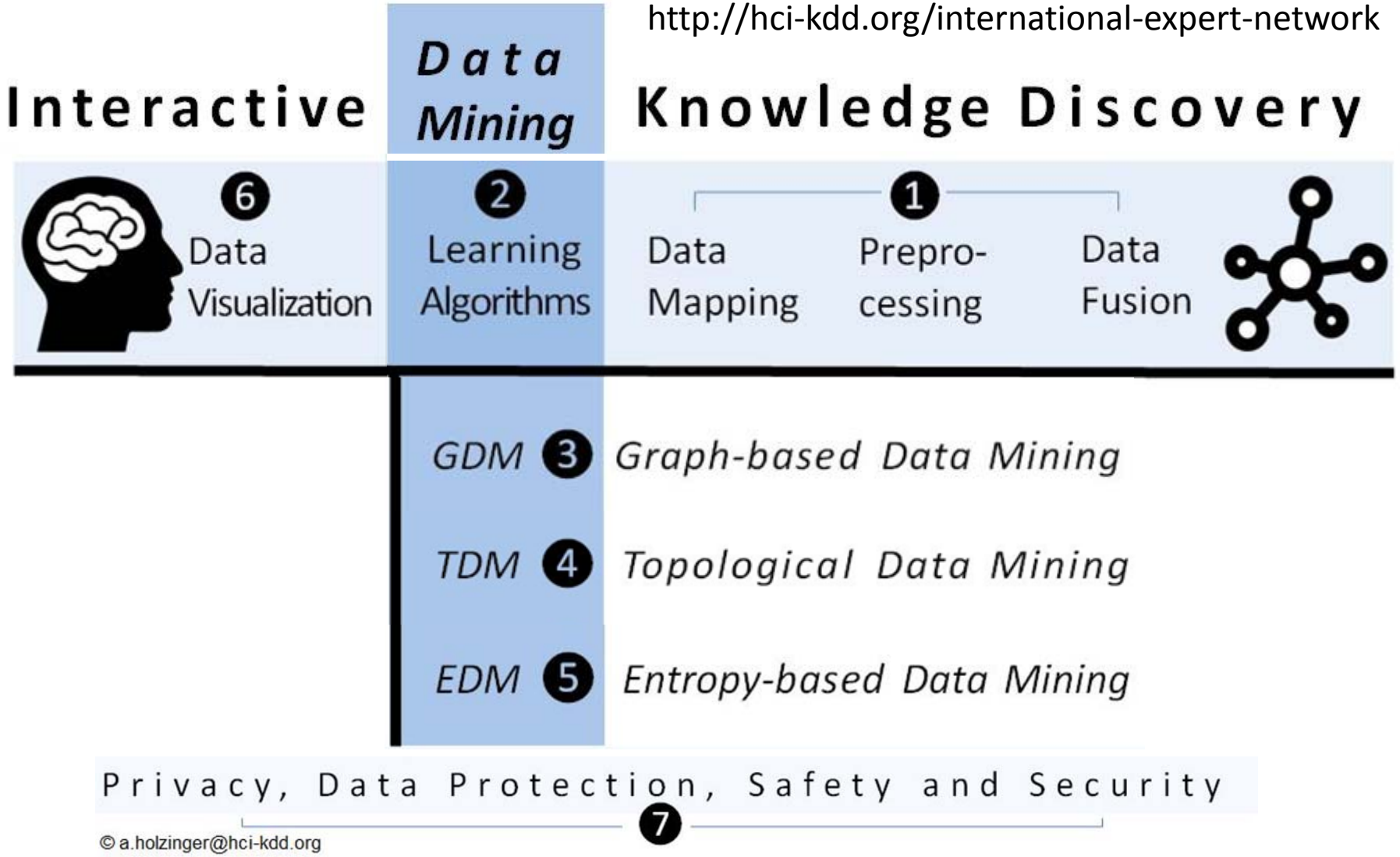
Introduction to word embeddings word-vectors (Word2Vec/GloVe) Tutorial

b.malle@hci-kdd.org

http://hci-kdd.org/machine-learning-for-health-informatics-course

ML needs a concerted effort fostering integrated research

http://hci-kdd.org/international-expert-network

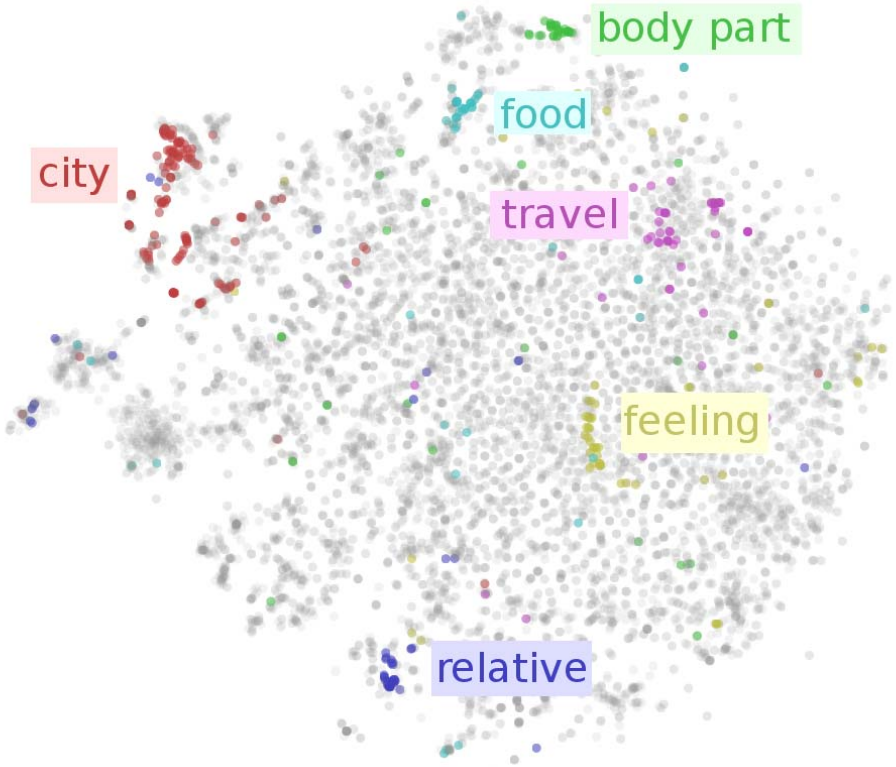


Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Overview of this Tutorial

1. What does Word embedding mean?
2. Three main ideas (take home message)
3. Word vector general assumptions
4. Advantages
5. Applications
6. Word vector operations
7. Methods (W2V, GloVe)
8. DEMO

Any technique mapping a word (or phrase) from it's original high-dimensional input space (the body of all words) to a lower-dimensional numerical vector space - so one *embeds* the word in a different space



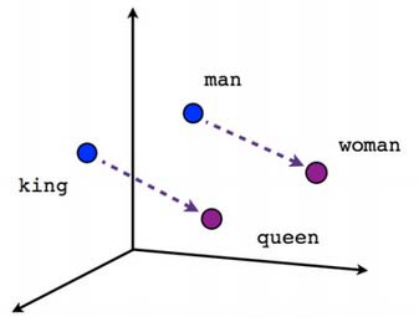
Source: http://sebastianruder.com/content/images/2016/04/word_embeddings_colah.png

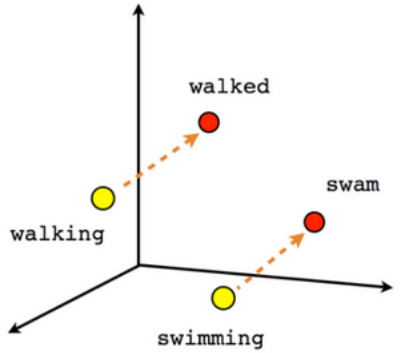
Points: original word space

Colored points / clusters: Word embedding

Learning word vectors for Sentiment Analysis [2]

- "Word representations are a critical component of many natural language processing systems. It is common to represent words as indices in a vocabulary, but this fails to capture the rich relational structure of the lexicon. Vector-based models do much better in this regard. They encode continuous similarities between words as distance or angle between word vectors in a high-dimensional space [1]".





https://www.tensorflow.org/versions/r0.7/tutorials/word2vec/index.html

[1] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. Learning word vectors for sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011. Association for Computational Linguistics, 142-150.

[2] Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V. & Holzinger, A. 2015. Computational approaches for mining user's opinions on the Web 2.0. Information Processing & Management, 51, (4), 510-519, doi:http://dx.doi.org/10.1016/j.ipm.2014.07.011.

Three main ideas (take-home messages)

1) Similarity in **meaning** ⟺ similarity in **vectors**

   ⇒ *Mathematics should be able to encode meaning*

2) You shall know a word by the company it keeps ;)

   ⇒ *The environment of a word gives meaning to it*

3) Use BIG datasets (millions of billions to words)

   ⇒ *Especially neural models require lots of data!*

Word Vectors General Assumptions

- The more often two words co-occur, the closer their vectors will be
- Two words have close meanings if their local neighborhoods are similar
- Maps of words (trained on the same dataset) should be similar for each language => can build translators!

Advantages

- Depending on the model, we do not infer meaning only from global statistics, but local contexts with are more precise

- Computers can "grasp" the meaning of words by looking at the distance between vectors / vector components

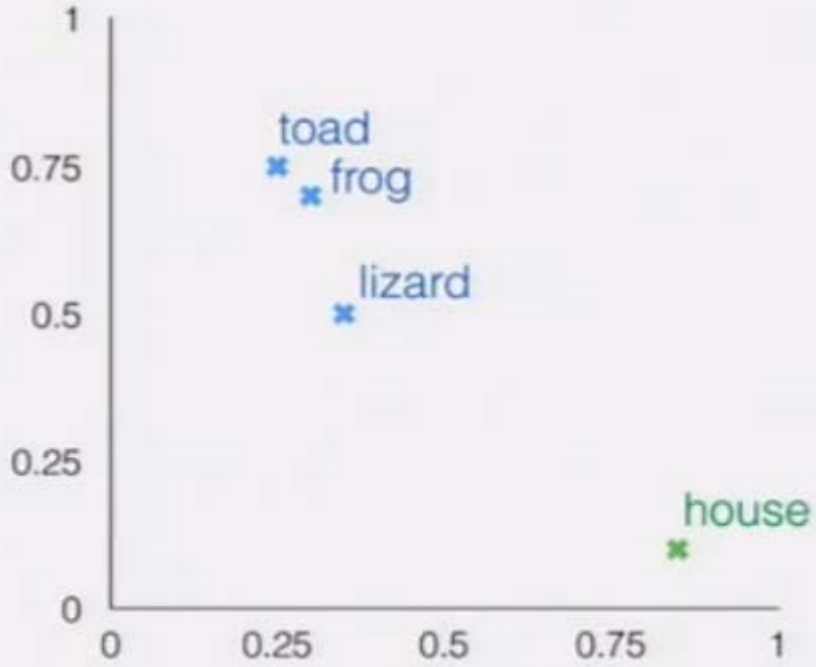
- Can apply mathematical operations translating to actual meaning within semantic models

Advantages

Distributed representations

Word vectors aren't guaranteed to encode any linguistic relationships between words, but many models produce vectors that do

frog	[0.30 0.70]
toad	[0.25 0.75]
lizard	[0.35 0.50]
house	[0.85 0.10]

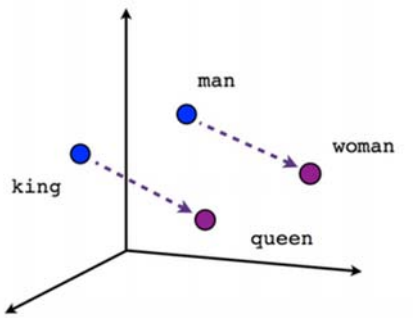


Source: https://www.youtube.com/watch?v=RyTpzZQrHCs

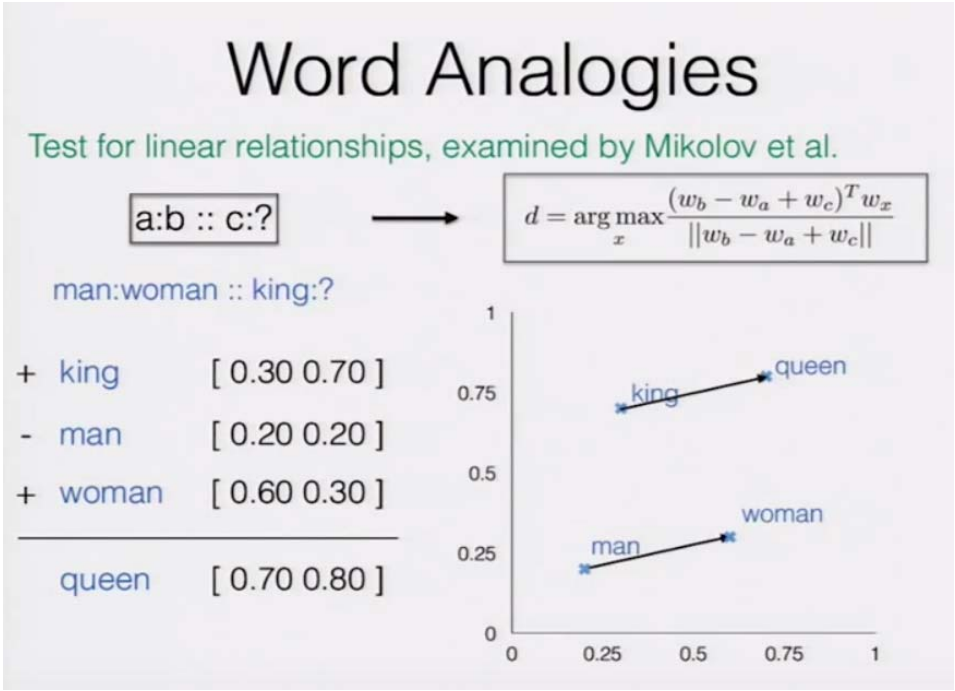
Applications

- Machine translation (word maps)
- Sentiment analysis
  - e.g. via Vector averaging
- Any NER task (seeking to locate and classify elements in text into pre-defined categories)
  - Topics
  - Goals
- Innovative search engine (ThisPlusThat), where we can subtract queryies, e.g.: "pizza + Japan - Italy => sushi“
- Thus: recommendation system for online shops
- ….

- Arithmetic



"king - man + woman = queen"

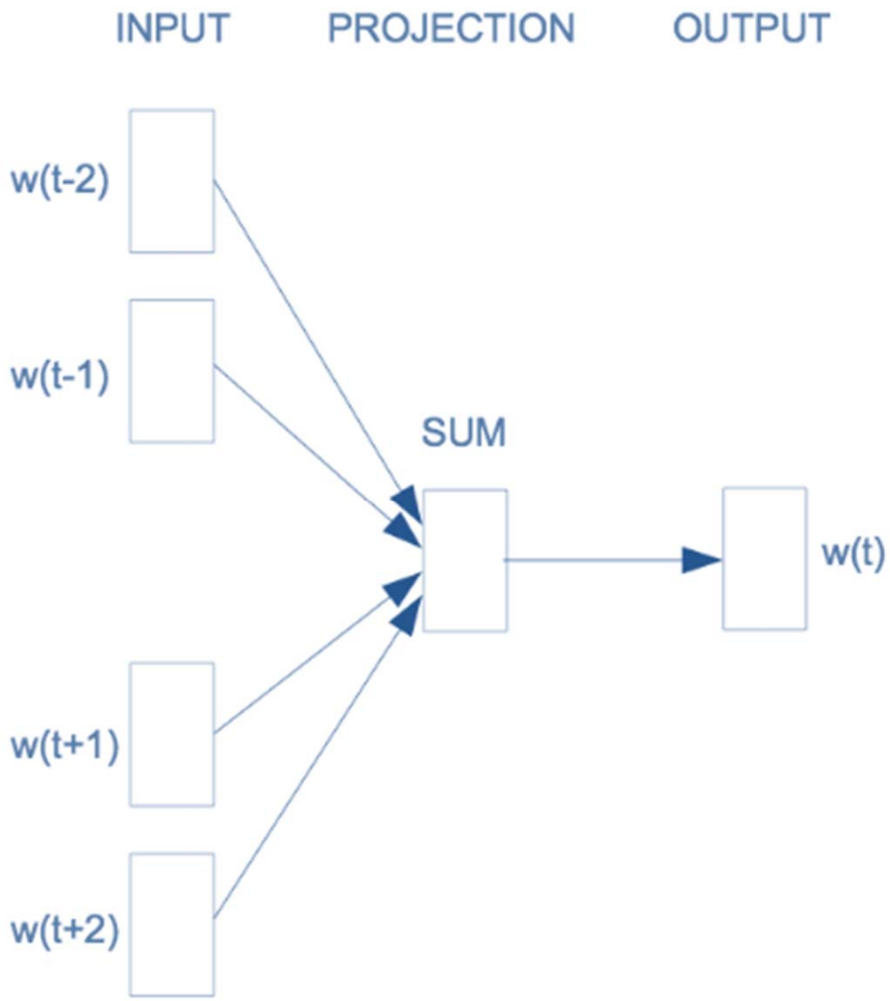


Source: https://www.youtube.com/watch?v=RyTpzZQrHCs

- 2. Nearest Neighbours
  - frog: toad, litoria, lizard, ...
  - works even with numbers (given certain context, like gene sequence)
- 3. Find words that do not belong
  - dog, cat, mouse, fruit basket

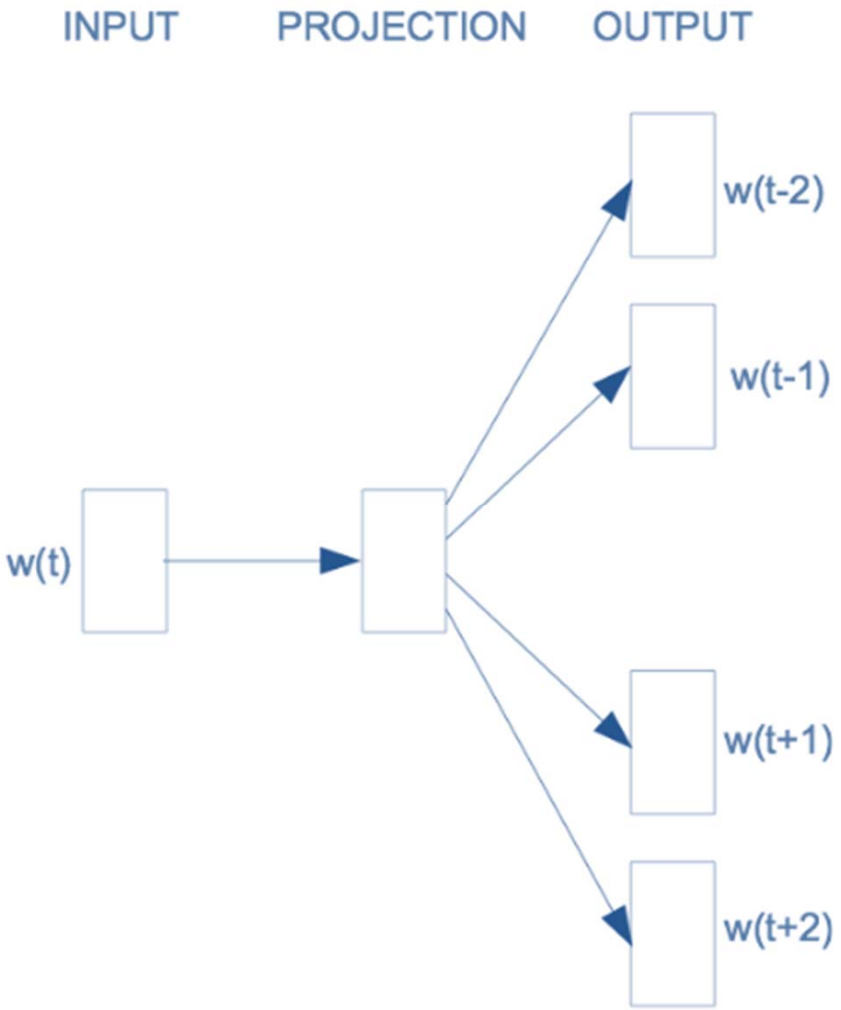
Methods – word2vec

- Uses neural networks to train word / context classifiers (feed-forward neural net)
- Uses local context windows (environment around any word in a corpus) as inputs to the NN
- Two main models:
  1. Continuous bag-of-words (CBOW)
  2. Skip-gram (SG)



CBOW uses the context given by a local window to predict a (known) center word

Skip-gram (SG) model



Skip-gram works exactly the other way around, using a given center word to predict it's context

Global Vectors for Word Embedding (GloVe)

1. Generate Co-occurrence matrix X (symmetric)

   - Take a context window (distance around a word, e.g. 10)
   - X(i,j) = # of times 2 words lie in the same context window

2. Factorize X

   - Extract vectors: X(i,j) = <vi,vj> | v el R^d (d is a hyperparamter)

3. Ratio of co-occurrences between 3 words => the ratio of the C/O of similar words will be closer to 1 => this is how meaning is captured in this model

Encoding meaning in vector differences

Crucial insight: Ratios of co-occurrence probabilities can encode meaning

	x = solid	x = gas	x = water	x = random
$P(x\|\text{ice})$	large	small	large	small
$P(x\|\text{steam})$	small	large	large	small
$\frac{P(x\|\text{ice})}{P(x\|\text{steam})}$	large	small	~1	~1

Source: https://www.youtube.com/watch?v=RyTpzZQrHCs

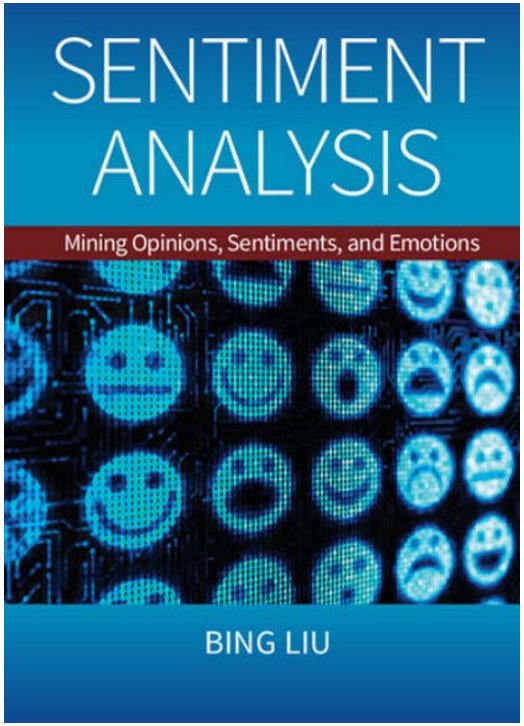
!!! DEMO !!!

Thank you!

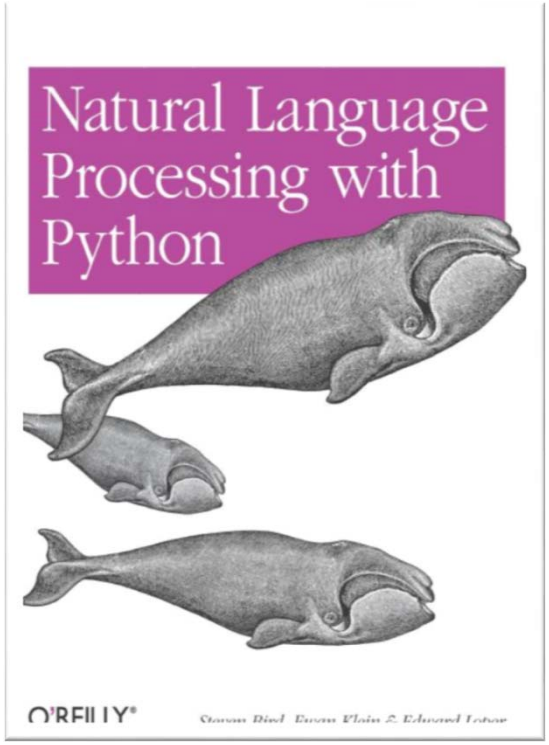
Sample Questions

- Principal question: Why is natural language important for health informatics?
- What are the key problems involved in NLP?
- Side Question: What makes NLP in the biomedical domain so difficult?
- What does word embedding mean?
- What are the three main ideas?
- Describe the advantages of this approach!
- Provide some examples for applications!

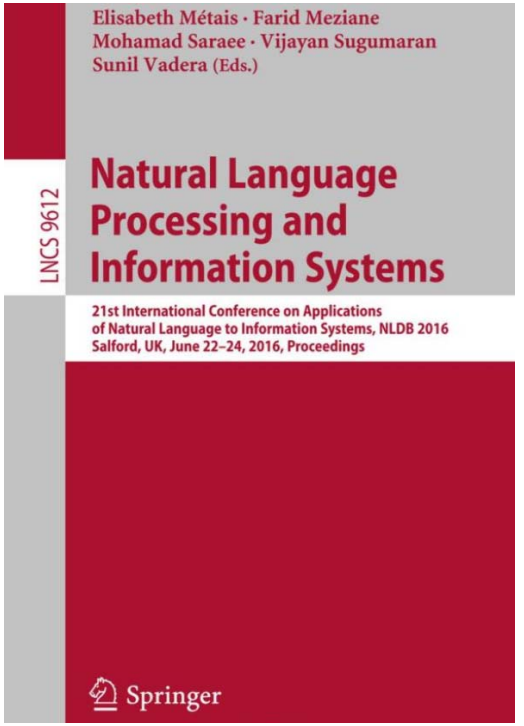
Recommended Reading



Liu, B. 2015. Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press.



Bird, S., Klein, E. & Loper, E. 2009. Natural language processing with Python, " O'Reilly Media



A Neural Probabilistic Language Model (1/2)

- 1. associate with each word in the vocabulary a distributed word feature vector - a real valued vector in $\mathbb{R}^m$
- 2. express the joint probability function of word sequences in terms of the feature vectors of these words in the sequence; and
- 3. learn simultaneously the word feature vectors and the parameters of that probability function

Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. 2003. A neural probabilistic language model. Journal of machine learning research (JMLR), 3, (2), 1137-1155.

A Neural Probabilistic Language Model (2/2)

Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. 2003. A neural probabilistic language model. Journal of machine learning research (JMLR), 3, (2), 1137-1155.