Lesson4:
Descriptive Modelling of Similarity of Text
Unit4:
Probabilistic (Language Models) based similarity measures

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

Completing this unit you should …

- Be aware of a unigram Language Model
- Know Laplacian (aka +1) smoothing
- Know the query likelihood model
- the Kullback Leibler Divergence
- See how a similarity measure can be derived from Kullback Leibler Divergence

Probability of a word to occur in a document

- Remember $D \subseteq W^*$

- For every Document D_i we get the maximum likelihood estimation by setting

$$P_{D_i}(w_j) = \frac{tf(w_j, D_i)}{len(D_i)}$$

- Where $len(D_i)$ returns the number of words in Document D_i

Removing zero Probabilities (+1 Smoothing)

- Set $\tilde{P}_{D_i}(w_j) = \dfrac{tf(w_j, D_i) + 1}{len(D_i) + N}$
- Where $N = |W|$ is the number of unique words in our corpus of documents
- Idea: increase every term frequency by 1
- Then normalize by N additionally seen words

This is what people do for simple querying and ranking

- Given a query $q = w_1 \ldots w_n$ of n words
- For all documents compute $r_1^{D_i} = \prod_{k=1}^{n} \tilde{P}_{D_i}(w_k)$
- Take the maximum over all documents

$$r_1 = \operatorname*{argmax}_{D_i \in D} \prod_{k=1}^{n} \tilde{P}_{D_i}(w_k)$$

- Which returns the document whose associated Model could most likely generate the query

Where are the similarities?!?

- This lesson is about using similarity measures for modeling and ranking
- Also this product is not a probability since it is not normed
- Is there a theoretically more beautiful way including similarities?

Roadmap (basically as in unit 1)

- Find a similarity measure to compare probability functions (associated to documents)
- Again: understand the query as a document
- Create its associated probability function as before
- Find the document whose probability function is most similar according to the measure

Kullback-Leibler Divergence is defined as

$$KL(D_i, D_j) = \sum_{w \in W} \tilde{P}_{D_i(w)} \log \frac{\tilde{P}_{D_i(w)}}{\tilde{P}_{D_j(w)}}$$

In information theory it is used as measure of difference (not distance) between two probability functions

BUT!!! $|KL(D_i, D_j)| \neq |KL(D_j, D_i)|$

No Symmetry!

Symmetrize the measure

$$SKL(D_i, D_j) := KL(D_i, D_j) + KL(D_j, D_i)$$

SKL is obviously symmetric

- Also $SKL(D_i, D_i) = 0 \forall D_i \in D$
  - Do you see why this equation holds?

- Remember we can derive a similarity via

$$\tilde{s}(D_i, D_j) = e^{-d(D_i, D_j)}$$

Open Question

Are the Query Likelihood model and the results from the similarity measure from the symmetrized Kullback-Leibler the same?

On empirical data it looks like that.

I could not prove this or find this in a paper.

Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: