Technische Universität München

sebis

Literature Study: Use Case Analysis for Word Embeddings

Bachelor Thesis – Kickoff Presentation

Nicolas Thule, 14.11.2016

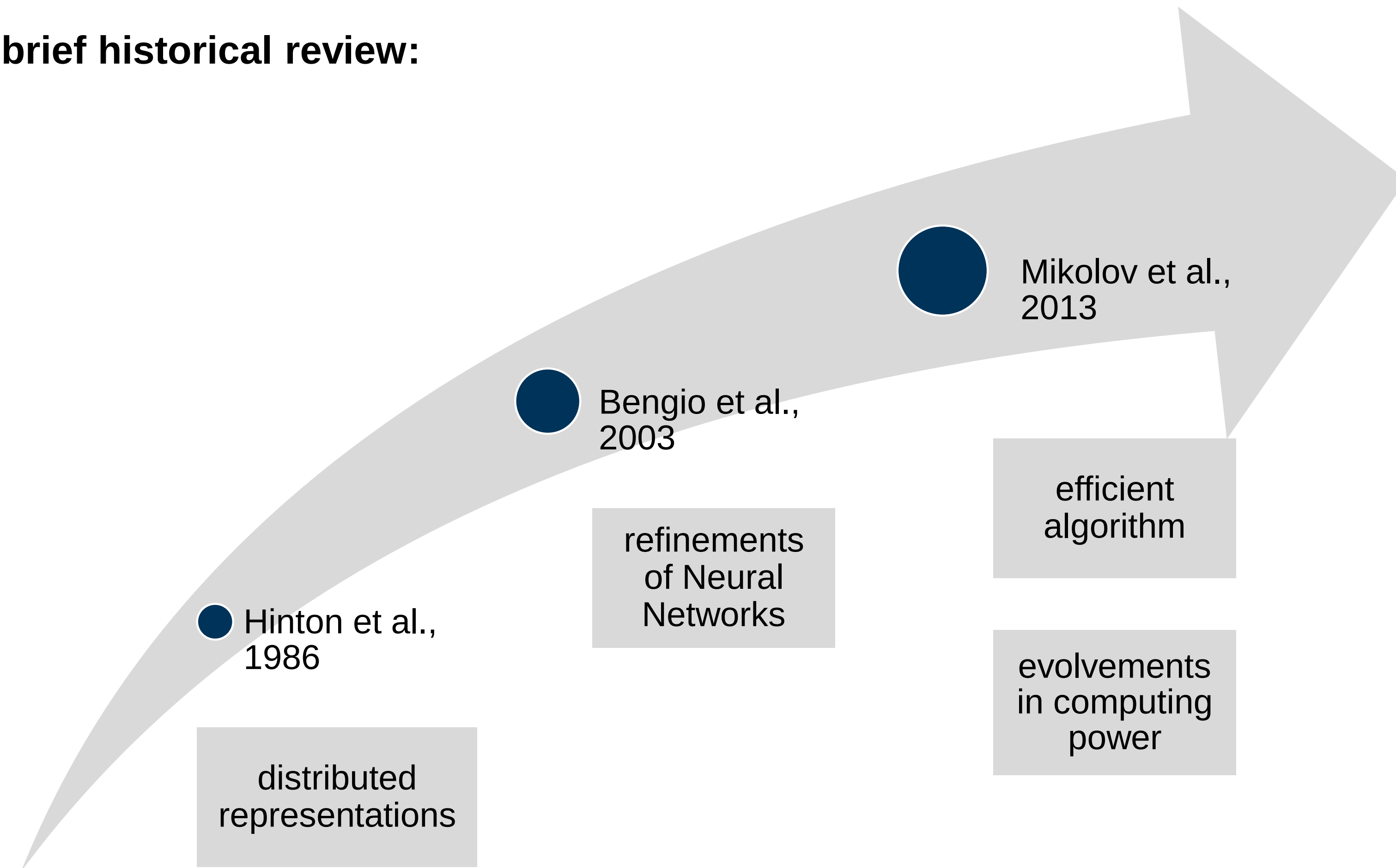
Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de

Table of Contents

sebis

A brief historical review:



Similar terms for word embeddings:
- distributional semantics
- distributed word representations
- semantic vectors

sebis

Traditional NLP

The best university in Germany.

Vocabulary

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Sparse one-hot representation

Vector length = size of vocabulary

Easy to use

Does not store context

Word Embeddings

The best university in Germany.

2,34	4,87	1.01	8,34	3,22
-1,30	3,22	0.01	0.23	5.67
0.23	-1.0	2.44	2.34	0.01
1.33	3.9	3.84	1.04	-1.2
1.0	3.8	-2.08	4,55	2,66

Manually chosen size

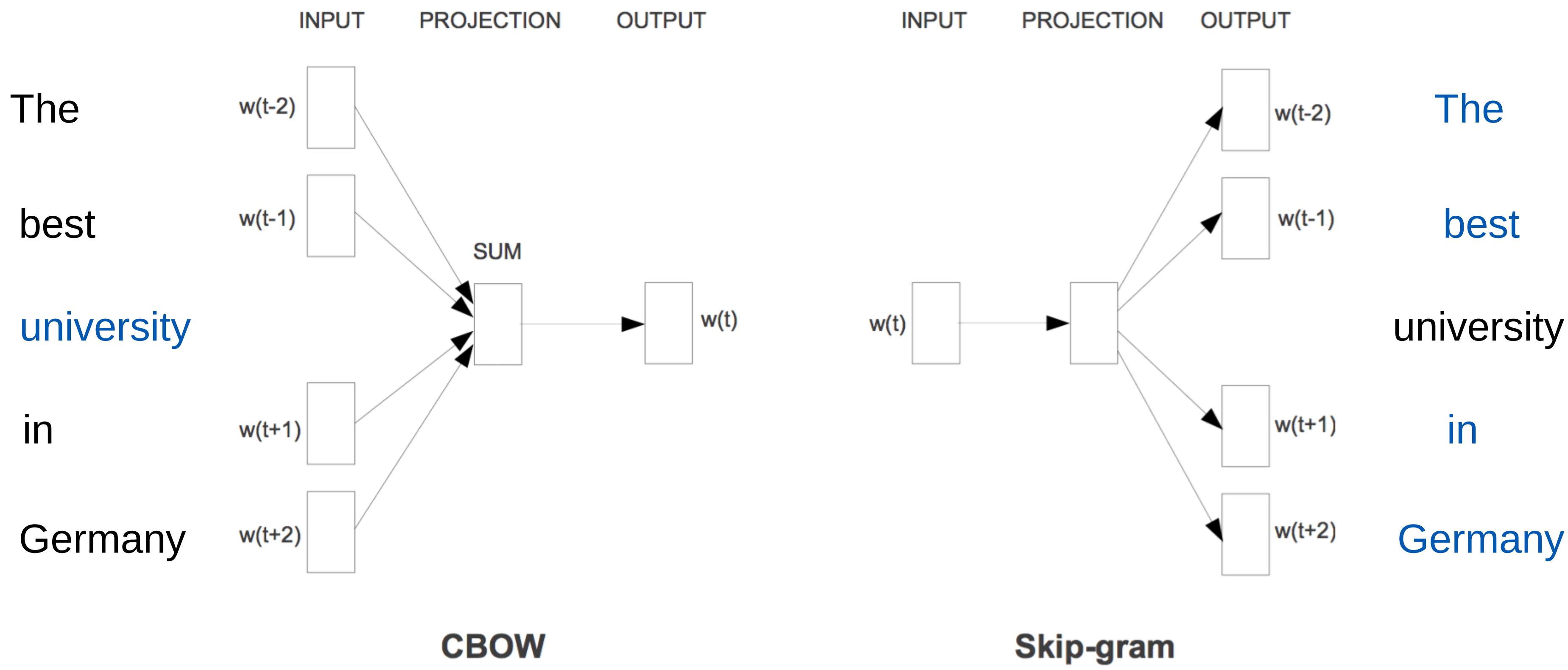
Dense representation

Every word represented as a vector of fixed length (e.g. 300)

Unsupervised learning

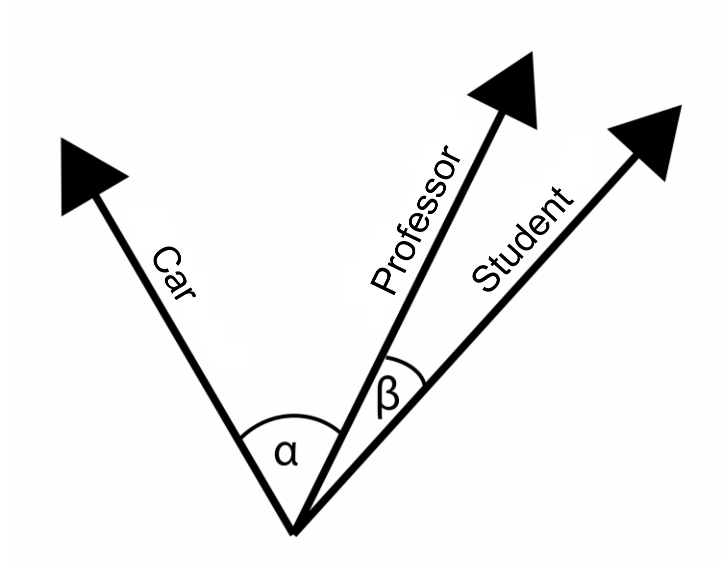
Frequent context of words is encoded in the vectors

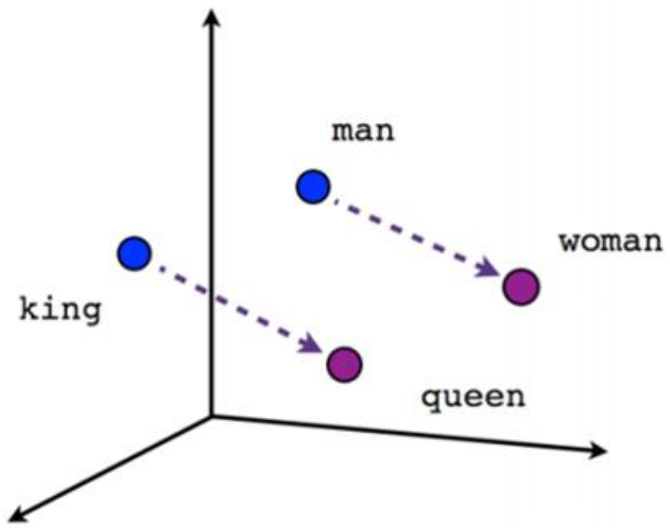
2 possible classification models with word2vec:



Mikolov et al., 2013

Illustration of the characteristics of Word Embeddings (in 2D/3D)





Simple mathematical operations, e.g. addition and substraction lead to interesting results:

Vec(„King“) – Vec („Man“) + Vec(„Woman“) -> Vec(„Queen“)

Mikolov et al., 2013b

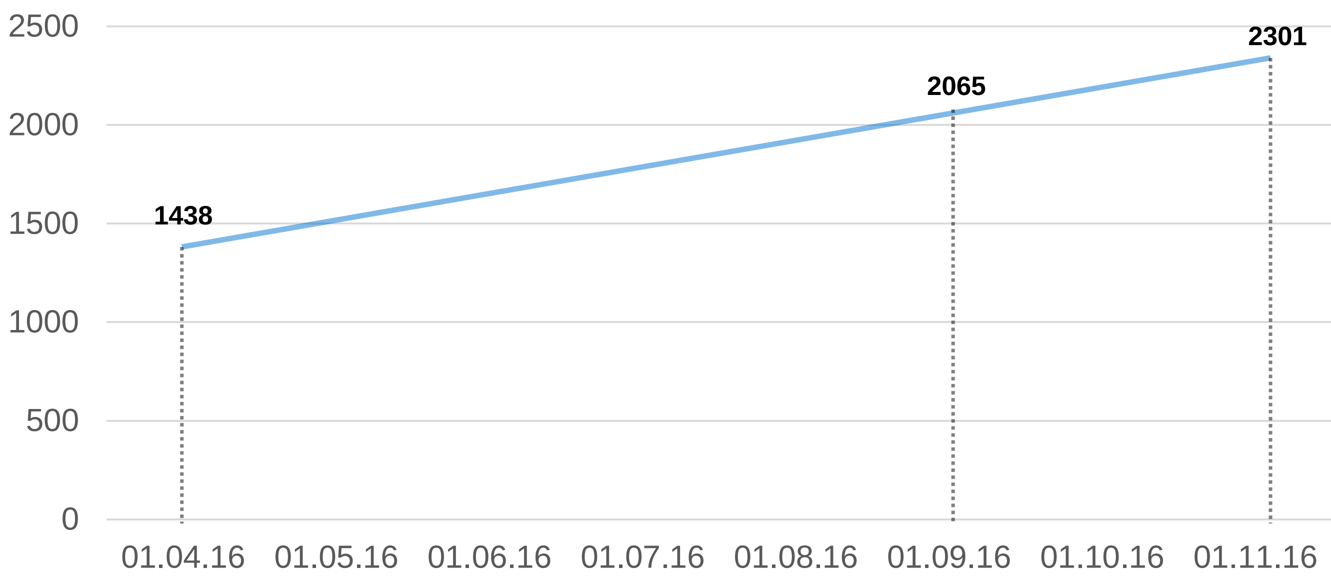
2. Motivation

Efficient estimation of word representations in vector space

T **Mikolov**, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org

Abstract: We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best ...

Zitiert von: 1438 Ähnliche Artikel Alle 8 Versionen Zitieren Speichern



=> no literature survey (esp. use cases) available

Table of Contents

sebis

3. Research Questions

?

Which Use Cases that apply to organizational contexts can be identified in the existing body of literature?

?

How can Use Cases for Word Embeddings be categorized?

?

What are the most frequent use cases?

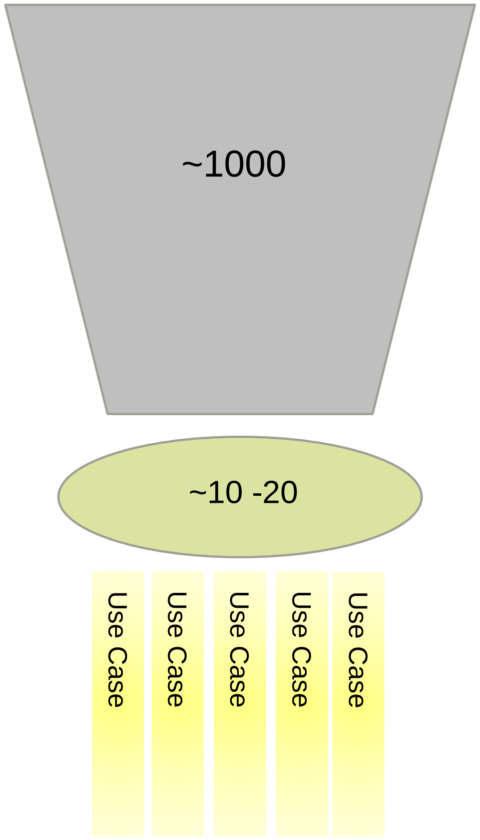
?

What further technological developments of Word2Vec can be identified in the body of literature?

?

For the most interesting Use Cases: How do they work technically?

4. Approach



1) Mikolov Cites Analysis

2) Use Case Selection
(in coordination with Prof. Matthes)

3) Focused Use Case Analysis

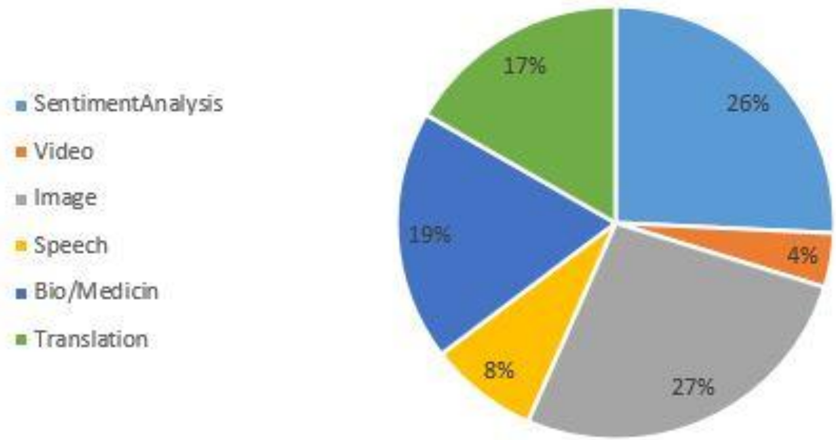
Table of Contents

5. Preliminary Results

sebis

Use Cases by Research Area:

- Sentiment Analysis
- Video: Annotation
- Image: Object Classification
- Speech Recognition
- Bio/Medicine:
  Genome Sequence Analysis
- Translation



Current baseline: 198 paper

1	title	authors	abstract	link	cites	categories	SubCategories
2	Distributed representations of words and phrases and the	T Mikolov, I Sutskever, K C	Abstract The	http://papers.	2271	Explanation	Other
3	Imagenet large scale visual recognition challenge	O Russakovsky, J Deng, I	Abstract The	http://link.spri	1268	UseCase	Image
4	Glove: Global Vectors for Word Representation.	R Socher, CD Manning	Theatre or th	http://llcao.ne	899	Advancement	Other
5	Distributed Representations of Sentences and Document	QV Le, T Mikolov	Abstract Man	http://www.jm	631	Advancement	Other
6	Show and tell: A neural image caption generator	O Vinyals, A Toshev, S Be	Abstract Auto	http://www.cv	500	UseCase	Image
7	Deep Learning	L Deng, D Yu	Abstract This	http://citeseer	303	Explanation	Other
8	Knowledge vault: A web-scale approach to probabilistic kr	X Dong, E Gabrilovich, G I	Abstract Rec	http://dl.acm.	270	UseCase	Other
9	Devise: A deep visual-semantic embedding model	A Frome, GS Corrado, J S	Abstract Mod	http://papers.	264	UseCase	Image
10	Intriguing properties of neural networks	C Szegedy, W Zaremba, I	Abstract: Dee	http://arxiv.or	236	UseCase	Image
11	Neural word embedding as implicit matrix factorization	O Levy, Y Goldberg	Abstract We	http://papers.	210	Explanation	Other
12	Exploiting similarities among languages for machine trans	T Mikolov, QV Le, I Sutske	Abstract: Dic	http://arxiv.or	206	UseCase	Translation
13	Unifying visual-semantic embeddings with multimodal neu	R Kiros, R Salakhutdinov,	Abstract: Insp	http://arxiv.or	174	UseCase	Image
14	Dependency-Based Word Embeddings.	O Levy, Y Goldberg	Abstract Whi	http://www.ac	169	Advancement	Other
15	Learning word embeddings efficiently with noise-contrasti	A Mnih, K Kavukcuoglu	Abstract Con	http://papers.	133	Advancement	Other
16	Bilingual Word Embeddings for Phrase-Based Machine T	WY Zou, R Socher, DM C	Abstract We	http://jan.stan	133	UseCase	Translation
17	Tailoring Continuous Word Representations for Depender	M Bansal, K Gimpel, K Liv	Abstract Wor	http://ttic.uchi	121	Advancement	Other
18	Multimodal Neural Language Models.	R Kiros, R Salakhutdinov,	Abstract We	http://www.jm	130	UseCase	Image

Benton et al., 2016

Vu et al., 2015

Ganguly et al., 2015;
Amer et al., 2016

Naderi and Hirst, 2015;
Rinott et al., 2015

Ozsow, 2016;
Barkan and
Koenigstein,2016

Iyyer et al., 2014

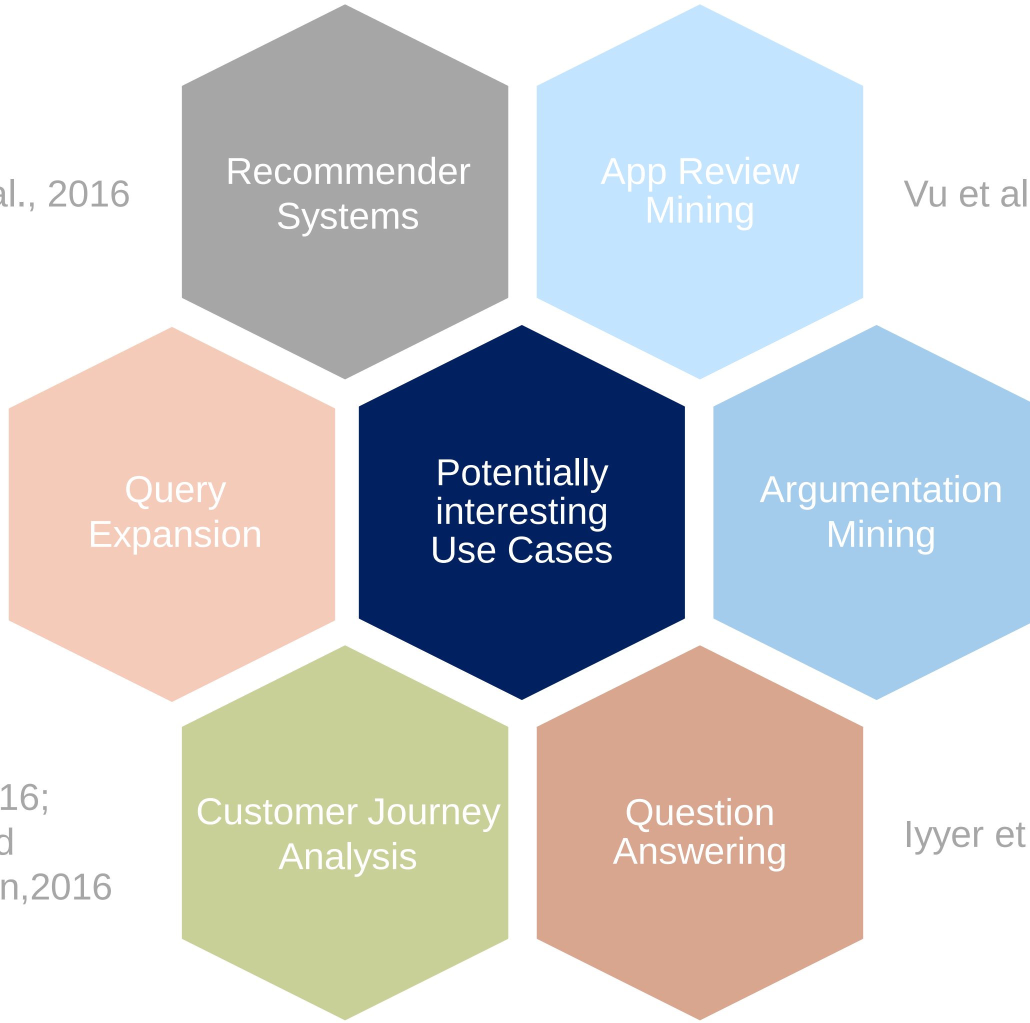


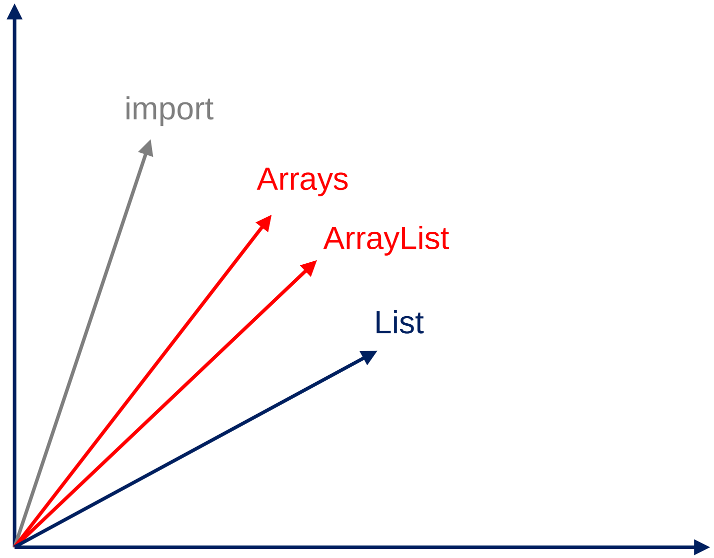
Table of Contents

sebis

Source Code Recommendation:

```
import java.util.Arrays;
import java.util.ArrayList;
import java.util.List;
```

```
[import, java, util, Arrays]
[import, java, util, ArrayList]
[import, java, util, List]
```



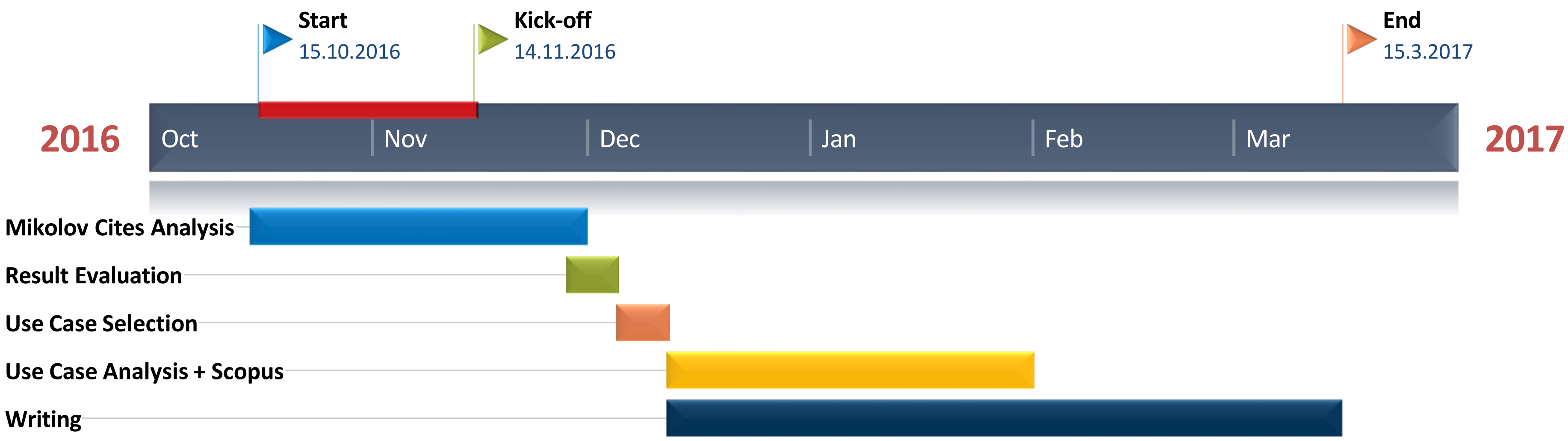

```
>>> closest_words('Array's, n=5)

[('Arrays'),
('equals()'),
('sort()'),
('toString()'),
('ArrayList')]
```

Gude, 2016

Table of Contents

sebis

7. Timeline



Start
15.10.2016
Kick-off
14.11.2016
End
15.3.2017
2016
Oct
Nov
Dec
Jan
Feb
Mar
2017
Mikolov Cites Analysis
Result Evaluation
Use Case Selection
Use Case Analysis + Scopus
Writing

Table of Contents

sebis

8. Sources

Recommended readings:

Natural Language Processing:

- COLLOBERT, Ronan, et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 2011, 12. Jg., Nr. Aug, S. 2493-2537.

Artificial Neural Networks:

- HAYKIN, Simon; NETWORK, Neural. A comprehensive foundation. *Neural Networks*, 2004, 2. Jg., Nr. 2004.
- SCHMIDHUBER, Jürgen. Deep learning in neural networks: An overview. *Neural Networks*, 2015, 61. Jg., S. 85-117.

Word Embeddings:

- PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. Glove: Global Vectors for Word Representation. In: *EMNLP*. 2014. S. 1532-43.

Word2Vec:

- MIKOLOV, Tomas, et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- https://deeplearning4j.org/word2vec
- https://rare-technologies.com/word2vec-tutorial/

8. Sources

- MIKOLOV, Tomas; LE, Quoc V.; SUTSKEVER, Ilya. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.
- BENGIO, Yoshua, et al. A neural probabilistic language model. *journal of machine learning research*, 2003, 3. Jg., Nr. Feb, S. 1137-1155.
- MELAMUD, Oren, et al. A simple word embedding model for lexical substitution. In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 2015. S. 1-7.
- https://gab41.lab41.org/python2vec-word-embeddings-for-source-code-3d14d030fe8f#.dw78fmh1h
- COLLOBERT, Ronan; WESTON, Jason. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th international conference on Machine learning*. ACM, 2008. S. 160-167.
- ALLAMANIS, Miltiadis, et al. Suggesting accurate method and class names. In: *Proceedings of the 2015* 10th Joint Meeting on Foundations of Software Engineering. ACM, 2015. S. 38-49.
- ALLAMANIS, Miltiadis; PENG, Hao; SUTTON, Charles. A Convolutional Attention Network for Extreme Summarization of Source Code. arXiv preprint arXiv:1602.03001, 2016.
- IYER, Srinivasan, et al. Summarizing Source Code using a Neural Attention Model. 2016
- https://www.tensorflow.org/versions/r0.11/tutorials/word2vec/index.html
- http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/
- HINTON, Geoffrey E.; MCCLELLAND, James L.; RUMELHART, David E. Distributed representations, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations. 1986.
- Gude A., Lab41, 2016: https://gab41.lab41.org/python2vec-word-embeddings-for-source-code-3d14d030fe8f#.dw78fmh1h

8. Sources

- VU, Phong Minh, et al. Mining User Opinions in Mobile App Reviews: A Keyword-Based Approach (T). In: *Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on*. IEEE, 2015. S. 749-759.
- NADERI, Nona; HIRST, Graeme. Argumentation Mining in Parliamentary Discourse. 2015
- RINOTT, Ruty, et al. Show Me Your Evidence–an Automatic Method for Context Dependent Evidence Detection. In: *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal*. 2015. S. 17-21.
- IYYER, Mohit, et al. A Neural Network for Factoid Question Answering over Paragraphs. In: *EMNLP*. 2014. S. 633-644.
- OZSOY, Makbule Gulcin. From Word Embeddings to Item Recommendation. *arXiv preprint arXiv:1601.01356*, 2016.
- BARKAN, Oren; KOENIGSTEIN, Noam. Item2Vec: Neural Item Embedding for Collaborative Filtering. *arXiv preprint arXiv:1603.04259*, 2016.
- GANGULY, Debasis, et al. Word embedding based generalized language model for information retrieval. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015. S. 795-798.
- AMER, Nawal Ould; MULHEM, Philippe; GÉRY, Mathias. Toward Word Embedding for Personalized Information Retrieval. In: *Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval*. 2016.
- ZAMANI, Hamed; CROFT, W. Bruce. Embedding-based query language models. In: *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*. ACM, 2016. S. 147-156.
- BENTON, Adrian; ARORA, Raman; DREDZE, Mark. Learning multiview embeddings of twitter users. In: *The 54th Annual Meeting of the Association for Computational Linguistics*. 2016. S. 14.

Technische Universität München

Thank you for your attention!
Questions?

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de