Semantic Representation of Documents

Corpus C: collection of documents
Vocabulary V: collection of terms

Vector Space Models (VSM)

Salton et al. (1975)

Assumption: document is a Bag of Words (BoW), i.e. only occurrences of terms in the text influence the meaning, neither grammatical dependencies, nor order of terms

Document is a $|V|$-dimensional vector with components representing the weight of each term in defining the meaning of text

Weighting Models:

- **Binary weights**
- **Term-Frequency**
- **Normalized Term-Frequency**
- **Term-Frequency and Inverse Document-Frequency**
- **Normalized Term-Frequency and Inverse Document-Frequency**

References: Manning et al. (1999, 2008); Jurafsky and Martin (2009)

Distributional Semantic Models (DSM)

Harris (1954); Firth (1957)

Assumption: *distributional hypothesis*, i.e. words with similar meaning appear in the same context

Count-based Models

Find a sub-space of the Vector Space Model (VSM) with enhanced ability in capturing word similarities
→Document is a K-dimensional vector with $K \ll |V|$

- **Latent Semantic Analysis (LSA)**
  Deerwester et al. (1990)
- **Non-negative Matrix Factorization (NMF)**
  Lee and Seung (1999)
- **Explicit Semantic Analysis (ESA)**
  Gabrilovich and Markovich (2007)

References: Manning et al. (2008); Jurafsky and Martin (2009); Aggarwal and Zhai (2013)

Probabilistic Topic Models

Document representation is the result of a stochastic generative process of words, based on hidden variables called *topics* that can be interpreted as themes discussed in text
→Document is a K-dimensional vector of proportions for each of the K topics

- **Probabilistic Latent Semantic Analysis (pLSA)**
  Hofmann (1999)
- **Latent Dirichlet Allocation (LDA)**
  Blei et al.(2003)
- **Pachinko Allocation Model (PAM)**
  Wei and McCallum (2006)

References: Blei (2012); Aggarwal and Zhai (2013)

Language Models

Assumption: documents are sequences of consecutive words
Goal: find the next word given a sequence of terms, defining word embedding that account for the context in which terms appear

Count-based Language Models

Goal: estimate the probability of observing a *target word* given its *context* through the *co-occurrence* of target and context words in the corpus
→ Define embedding for words

Examples: Positive Pointwise Mutual Information (PPMI), N-grams

References: Manning et al. (1999); Bullinaria and Levy (2007); Jurafsky and Martin (2009); Turney and Pantel (2010)

Predictive Language Models
a.k.a.
Neural Language Models

Goal: estimate the probability of observing a *target word* given its *context learning a language model* on a corpus of documents and using such model to *predict* the probability of observing a new word

Skip-gram and Continuous Bag Of Words (CBOW) Models
a.k.a.
"word2vec"
Mikolov et al. (2013a,b,c)
→ Define embedding for words

References: Mikolov et al. (2013a,b,c)
Suggested Readings: Baroni et al. (2014); Levy et al. (2015); Goldberg and Levy (2014); Caselles-Dupré (2015)

Paragraph Vector Models
a.k.a.
"doc2vec"
Le and Mikolov (2014)

Enrich the word2vec architectures learning also an embedding vector for the chunk of text, called *paragraph*, from which words have been extracted
→ Document, particular case of *paragraph*, is a K-dimensional vector

References: Le and Mikolov (2014)
Suggested Readings: Lenci (2018)

References

- *Aggarwal and Zhai (2013):* Aggarwal, C. and Zhai, C. (2013). An introduction to text mining.
- *Baroni et al. (2014):* Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. volume 1, pages 238–247.
- *Blei (2012):* Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4):77–84.
- *Blei et al. (2003):* Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(4-5):993–1022.
- *Bullinaria and Levy (2007):* Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior research methods, 39(3):510–526.
- *Caselles-Dupré (2015):* Caselles-Dupré, H., Lesaint, F., and Royo-Letelier, J. (2018). Word2vec applied to recommendation: Hyperparameters matter. arXiv preprint arXiv:1804.04212.
- *Deerwester et al. (1990):* Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407.
- *Gabrilovich and Markovich (2007):* Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. pages 1606–1611.
- *Goldberg and Levy (2014):* Goldberg, Y. and Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.
- *Hofmann (1999):* Hofmann, T. (1999). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pages 289–296. Morgan Kaufmann Publishers Inc.
- *Jurafsky and Martin (2009):* Jurafsky, D. and Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ: Prentice Hall.
- *Le and Mikolov (2014):* Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. volume 4, pages 2931–2939.
- *Lee and Seung (1999):* Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791.
- *Lenci (2018):* Lenci, A. (2018). Distributional models of word meaning. Annual Review of Linguistics, (4):151–171.
- *Levy et al. (2015):* Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3:211–225.
- *Manning et al. (1999):* Manning, C. D. and Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- *Manning et al. (2008):* Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to Information Retrieval. New York: Cambridge University Press.
- *Mikolov et al. (2013a):* Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- *Mikolov et al. (2013b):* Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.
- *Mikolov et al. (2013c):* Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations ofwords and phrases and their compositionality.
- *Salton et al. (1975):* Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620.
- *Turney and Pantel (2010):* Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37:141–188.
- *Wei and McCallum (2006):* Wei, L. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. volume 148, pages 577–584.