Task 2: Semantic similarity of words

Data

SimLex999 [1]: https://fh295.github.io/simlex.html

Methods

1. WordNet https://wordnet.princeton.edu/
   WordNet-based similarity in NLTK: https://www.nltk.org/howto/wordnet.html#similarity
2. fastText embeddings [2, 3] https://fasttext.cc/
   Multilingual models (bottom of the page): https://fasttext.cc/docs/en/crawl-vectors.html
   Python module: https://fasttext.cc/docs/en/python-module.html

Subtasks and points

1. Install NLTK, download WordNet data.
2. Download and review *SimLex999* data.
3. Calculate word similarities based on WordNet's `path_similarity` (iterate over all synsets pairs the words belong to, account for POS tags). Are any words from SimLex999 missing in WordNet?
4. Install fastText, download English fastText model in binary format (https://fasttext.cc/docs/en/crawl-vectors.html).
5. Calculate word similarities based on cosine similarity of word vectors (note that e.g. `scipy.spatial.distance.cosine` returns). Report if any words are missing in the model.
6. Calculate Kendall's tau (e.g. using `scipy.stats.kendalltau`) between the gold standard and obtained scores (use only word pairs processed by all models). Summarize findings in a table and analyze them.

References

1. Hill, Felix, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." *Computational Linguistics* 41.4 (2015): 665-695.
2. Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information." *Transactions of the association for computational linguistics* 5 (2017): 135-146.
3. Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. "Advances in pre-training distributed word representations." *arXiv preprint arXiv:1712.09405* (2017).