Computer-Linguistische Anwendungen

CLA | B.Sc. | LMU

FastText

FastText

- FastText ist eine Erweiterung des Skipgram word2vec
- Es berechnet zusätzlich die Embeddings für Character-NGrams
- Das Embedding eines Wortes ist dann: Die Summe der gewichteten Character NGram Embeddings
- Parameter:
  - Minimum Ngram Länge: 3
  - Maximum Ngram Länge: 6
- Das Embedding von “Dendrite” wird demnach die Summe der folgenden NGrams:
  - (1gram) : @dendrite@
  - (3gram) : @de den end ndr dri rit ite te@
  - (4gram) : @den dend endr ndri drit rite ite@
  - (5gram) : @dend dendr endri ndrit drite rite@
  - (6gram) : @dendr dendri endrit ndrite drite@

Also insgesamt: @dendrite@ @de den end ndr dri rit ite te@ @den dend endr ndri drit rite ite@ @dend dendr endri ndrit drite rite@ @dendr dendri endrit ndrite drite@

FastText

- Beispiel 1: Embeddings für Character-Ngram “dendri”
  - “dendrite” und “dendritic” sind ähnlich

- Beispiel 2: Embeddings für Character-Ngram “tech-”
  - “Tech-rich” und “tech-heavy” sind ähnlich

Buchstaben n-Gram Generalisierung kann **gut** sein

word2vec

1.000 automobile 779 mid-size 770 armored 763 seaplane 754 bus 754 jet 751 submarine 750 aerial 744 improvised 741 anti-aircraft

FastText

1.000 automobile 976 automobiles 929 Automobile 858 manufacturing 853 motorcycles 849 Manufacturing 848 motorcycle 841 automotive 814 manufacturer 811 manufacture

Buchstaben n-Gram Generalisierung kann **schlecht** sein

word2vec

1.000 Steelers 884 Expos 865 Cubs 848 Broncos 831 Dinneen 831 Dolphins 827 Pirates 826 Copley 818 Dodgers 814 Raiders

FastText

1.000 Steelers 893 49ers 883 Steele 876 Rodgers 857 Colts 852 Oilers 851 Dodgers 849 Chalmers 849 Raiders 844 Coach

Buchstaben n-Gram Generalisierung

word2vec

("video-conferences" did not occur in corpus)

FastText

1.000 video-conferences 942 conferences 872 conference 870 Conferences 823 inferences 806 Questions 805 sponsorship 800 References 797 participates 796 affiliations

Further readings

Word2vec:

https://code.google.com/archive/p/word2vec/

FastText:

https://fasttext.cc/docs/en/support.html

TensorFlow:

https://www.tensorflow.org/text/guide/word_embeddings

Zusammenfassung

Zusammenfassung

- Drei Versionen des word2vec Skipgram
    - Matrix Faktorisierung (SVD) der PPMI Matrix
    - Skipgram Negative Sampling (SGNS) mit Gradient Descent
    - Hierarchical Softmax

Zusammenfassung

- Embeddings mit Gradient Descent lernen:
  - Kostenfunktion in Negative Sampling:
    - Mache das dot product der “true” Paare so groß wie möglich und das er “false” Paare so klein wie möglich
  - Anzahl der Parameter: $2d\,|V|$
  - Gradient Descent

Zusammenfassung

- Visualisierung
  - Oft verwendet: zweidimensionale Projektion
    - PCA (Principal Component Analysis)
    - t-SNE (t-distributed stochastic neighbor embedding)
    - Gensim library
  - Keine direkte Interpretation der kleineren Dimensionen möglich

Zusammenfassung

- FastText
  - Lernt Embeddings für Character ngrams
  - Kann mit out-of-vocabulary (OOV) Wörtern umgehen
  - FastText gibt (“automobile”), FastText nimmt (“Steelers”)