Conventional Approaches - takeaway notes

Yan Song

TF/IDF

- Why IDF is important?
  - Filter out meaningless high-frequent terms
  - Emphasize the key terms w.r.t. docs
- Effeciency?
  - No if there are huge numbers of docs (terms)
- Sparseness?
  - Yes, most of them will be zero
- Quality?
  - Depending on the unit that who the terms are clustered (doc)

LSA

- Why using SVD or similar tricks?
  - High-dimension to low-dimension
  - Frequency-based measurement to dense representation
- Efficiency?
  - No if there are huge number of docs (terms) SVD can hardly be paralleled.
- Sparseness?
  - No, dense representation
- Quality?
  - Depending on doc, no particular meaning on each dimension

LDA

- Why using Gibbs sampling?
  - An alternative of matrix factorization
  - Frequency-based measurement to dense representation
- Efficiency?
  - OK, but takes large memory to store two huge matrices
- Sparseness?
  - No, dense representation
- Quality?
  - Depending on doc, no negative values (probabilities)

Summary

- Bag-of-words representations
  - No structure information embedded (e.g., ordering of words)
- Frequency based
  - Reflects the strength of a term using documents, or vice versa
- Requires another reference system
  - Requires another axis (e.g., document, topic) to represent words
- Static
  - Requires offline training, restricted in generalization

Hw2

- Understand GloVe: https://nlp.stanford.edu/pubs/glove.pdf
- Train your GloVe embedding using enwiki8 corpus.
  - GloVe: https://nlp.stanford.edu/projects/glove/
  - enwiki8: http://mattmahoney.net/dc/enwik8.zip
- Settings
  - 200d
  - +/-5 word window
  - frequency cut-off: 5
- Evaluate your embeddings by scoring word pairs
  - Coseine similarity