Word Vectors and FNC

AI4ALL NLP Group

Overview of the Week

- Wednesday
  - 9:30-12: Review + time to work on presentations.
  - 1-3: Word vectors with FNC + time to work on presentations.
- Thursday
  - 9:30-12: Putting it all together + time to work on presentations.
  - 1-3: Minipresentations (1-2:30 per person) + time to work on presentation.
- Friday
  - 9:30-10:30: Free time to finish/practice presentation.
  - 10:30-12, 1-3: Presentations with other groups.

Wrapping up FNC

Today and tomorrow: shorter FNC slides+workshops, more time to work on presentations.

Today: Short slides+exercises with semantic similarity, then time to work on presentations.

A set of pre-trained word vectors

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

word2vec

https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
https://code.google.com/archive/p/word2vec/

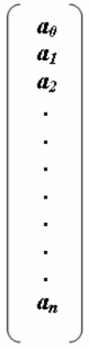
Converting a headline into a vector

How can we go from

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

to



https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
https://www.researchgate.net/figure/Fig-2-Graphical-illustration-of-matrix-converter-operations-for-both-1D-vector-to-2D_fig2_266674755

Once we have vectors for headlines and article bodies, what next?

Quick workshop, then presentation work-time