Extracting keywords from emails using distributed word vectors

Irving Rodriguez

Stanford University

Keywords

- Currently, keyword extraction relies on statistical models, like **term frequency-inverse document frequency (TFIDF)**
- However, these models do not use any of the linguistic or syntactic information contained in the content of the email.
  - ➢ **Is it possible to extract better keywords using the relationships between the words of the email?**
    - ➢ **Yields better information!**
  - ➢ Solution explored: **Distributed Word Vectors**
- Can this approach handle other challenges?
  - Word disambiguation (“apple”: fruit or company?)
  - Semantic analysis
  - Generality of keywords

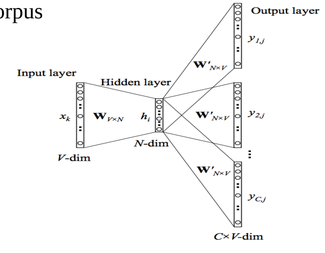
Why Word Vectors?

- Assumption: words with similar linguistic/syntactic characteristics appear in similar **contexts** ⟶ similar words will have similar vectors
- Word vectors have interesting observed properties in literature!

- Train vectors using **continuous skip-gram** algorithm:
  - Maximize log-probability of surrounding context words given a center word:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t) \quad \text{, with} \quad p(w_O|w_I)=\frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_{w}}^{\top} v_{w_I}\right)}$$

- Use neural network with projection layer and softmax layer (above) trained on many contexts of a corpus



- Can omit common words (“the”, “and”) by omitting words with probability depending on word frequency in corpus and **subsampling threshold** t:

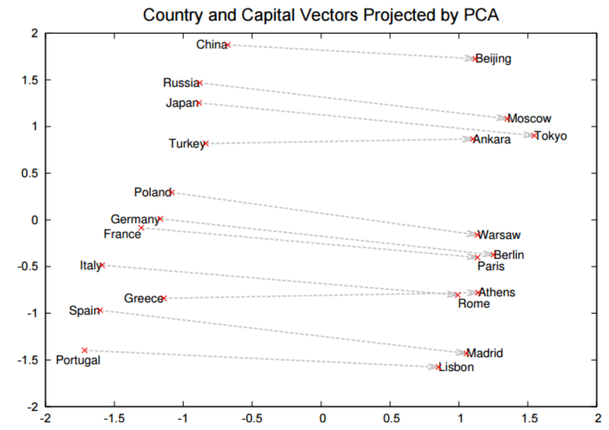
$$P(w_i)=1-\sqrt{\frac{t}{f(w_i)}}$$

Training and Testing of Vectors

- **Training** model using subset of English Wikipedia (~2M articles, ~675M tokens), using publicly availably `word2vec` implementation from Google

- **Testing:** test model’s ability to capture word relationships via **vector algebra:**

1) Present model with 4-word string, i.e. Chicago Illinois Detroit Michigan
2) If, i.e. v(Chicago) – v(Illinois) + v(Detroit) outputs v(Michigan) using **cosine similarity,**
3) test output = 1. Else 0.



Visualized example of implicitly learned word relationships using word vectors due to similar contexts between related words. ***Source: Mikolov et. al****, “Disributed Representations of Words and Phrases and their Compositionality”*

Final Word Vector Model Parameters

Vocabulary Size: 41k

Dimension	Context Window Size	Minimum Token Appearance	Subsampling Threshold
300	12	50	1e-5

Word Vectors on Email Data

- Use TF-IDF to identify “keywords” from 100 personal emails
  - Use word vector model on individual email keywords

Hope this email finds you well. Attached you will find the teams for the winter season for your chapter. Right now, if you look in your team finder report, it only shows you fall teams. I will contact IT to make sure your winter team is showing.

Once she updates your team finder report, you will be able to create this type of excel sheet for the spring. All you would need to do is, when you are in the team finder report, press printable view, then open in excel.

From there you can delete columns and rows of info you don’t need, as well as shrink them to your preference. If you have any questions please let me know. Oh also please be aware that we have some teams with the last date to apply is on 11/20, so you might want to try to reach out asap if not, please delete those teams and move forward with the teams with a later last date to apply.

Thanks

Kevin Reduta
Regional Program Coordinator, Northern California

Coaching Corps
310 Eighth Street, Suite 300
Oakland, CA 94607
(510) 496-5129
www.CoachingCorps.org

Top TF-IDF Keywords	Similar Words from Word Vector Model
coaching	player, football, manager, team, basketball
season	post-season, playoffs, team, league,scoring
team	championship, win, teammates, league, season
California	Pasadena, Sacramento, Monterey
winter	summer, autumn, snowfall, spring

- Combining keywords yields more meaningful information, more closely related to content:

Keyword Pairs	Similar Words from Word Vector Model
team, coach	championship, basketball, undefeated, captain, sports
winter, team	season, league, playoffs, game, sports

Discussion

- Generally, the word vector model appears to extract more information using its background knowledge of word contexts
  - Need more rigorous way of showing “more information”
- Word vectors do not require entire “email” corpus to calculate keywords; can be complementary to TFIDF
- Conceptually, can capture more complex relationships in content

Future Work

- Use more complete corpus to train word vector model (including phrase-identification in corpus and emails)
- Quantitatively explore linguistic relationships: Some Linguistics: Paradigmatic

John bought a new, expensive ***car***

Synonymy: *purchased acquired* *Hypernymy:* *vehicle machine*

Antonymy: *sold*

References

- Bengio et al. “A Neural Probabilistic Language Model”. http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf
- Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. http://arxiv.org/pdf/1301.3781.pdf
- Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf