Text Mining Lab

Summer 2017

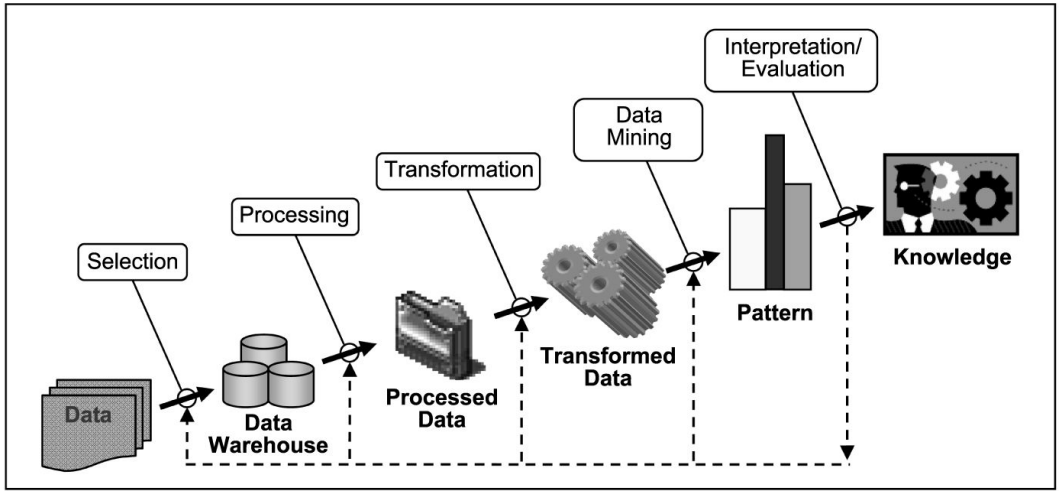
Elvis Saravia
PhD, Information Systems and Applications
ellfae@gmail.com
Github username: omarsar

Questions: sli.do (#Z217)

King - man + woman = ?

Expectations for this lab

- Environment Setup
- Data preprocessing
- Training Models
- Evaluation of Models
- Assignments (best part)
- Project (3 options)
- To infinity and beyond (closing remarks)



Knowledge Discovery (KDD) Process

Word Vector Representations

Represent the meaning of a word?

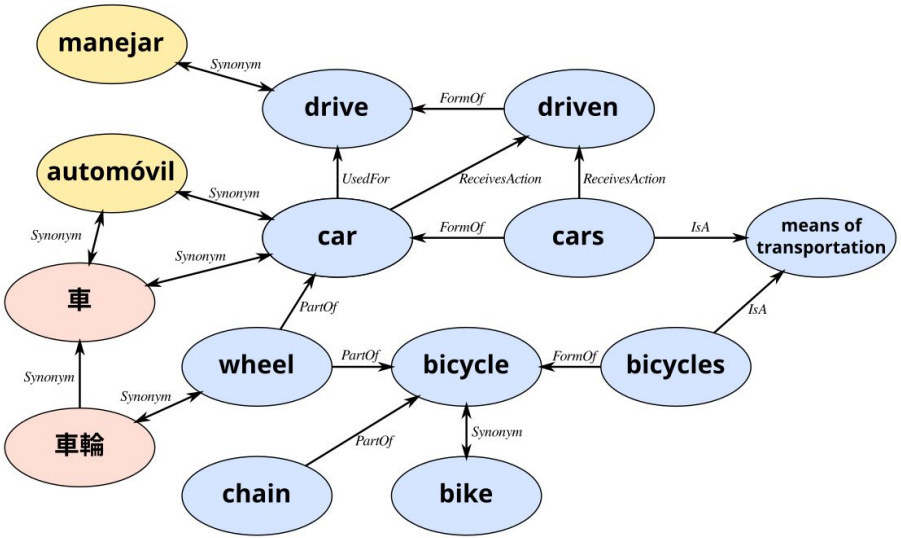
Words and phrases *directly represent* an idea

Words and signs are used *to express* an idea in work of writing, art, etc.

How does a computer represent meaning of a word?

Represent the meaning of a word on a computer?

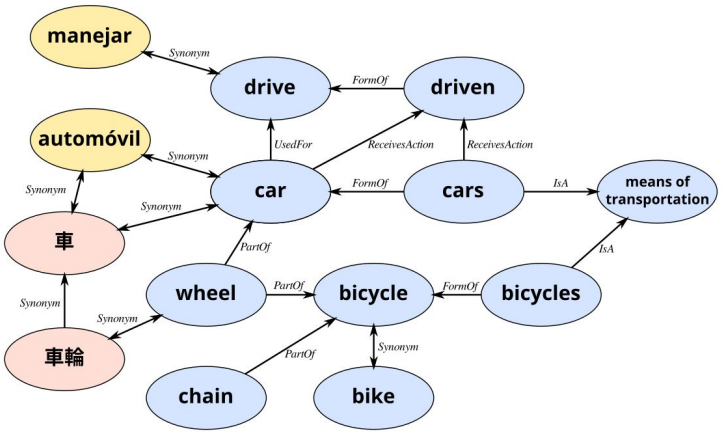
Solution: Taxonomy, such as WordNet and ConceptNet, that contains hypernyms (is-a) relationships and synonyms sets.



ConceptNet

Problems with Discrete Representation

- **Low Coverage** - fails to capture all word nuances (e.g., synonyms)
- **Difficult to keep up to date** - we just keep inventing new words like *boo* and *fab*
- **Subjective** - because it requires human annotation



Problems with Discrete Representation

Most Natural Language Processing (NLP) and rule-based approaches regard words as **atomic symbols** (“*each word a nation on its own*”)

- **Word Similarity Fails** - no clear *relationship* between words
- **Curse of Dimensionality -** too many dimensions; too much sparsity; memory inefficient

One-hot representation

Motel = [0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0]
Hotel = [0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0]

$$\vec{Motel} \cdot \vec{Hotel}^T = 0$$

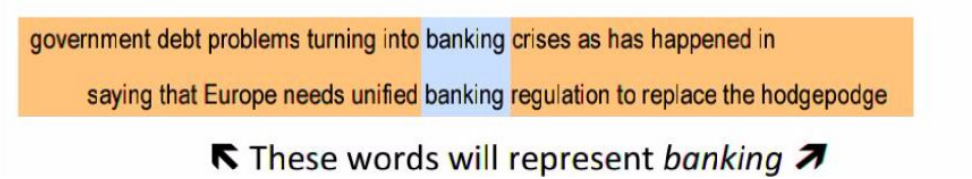
Distribution Similarity Based Representations

Idea: represent words through it neighbours or the context in which they are used

Solution: dense vector representation for predicting words appearing in its context

"You shall know a word by the company it keeps"

-J. R. Firth 1957



Distributed representation (low-dimension vector)

hotel = [0.728 0.234 -0.23 0.223]

Word2vec (faster and simpler)

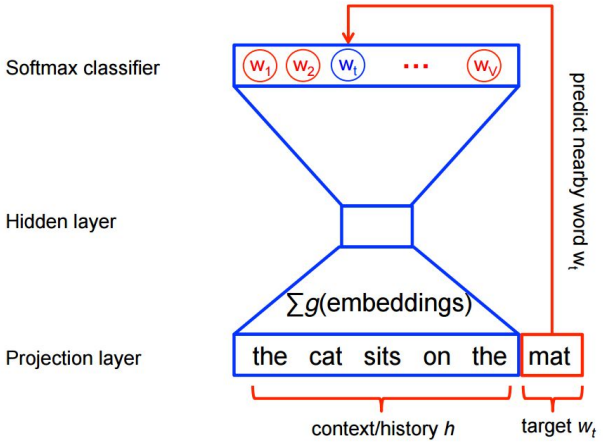
Ideas:

1. Word vectors are trained so that they become **good features** for predicting context (surrounding) words
2. Every word is mapped to a **unique word vector**
3. Similar words tend to be **close to each other** in a vector space

Algorithm:

1. Initialize random vectors
2. Pick an objective function
3. Do gradient descent

Paper source: https://arxiv.org/pdf/1301.3781.pdf



Architectures: CBOW and Skip-gram

CBOW - predicts the current word based on the context

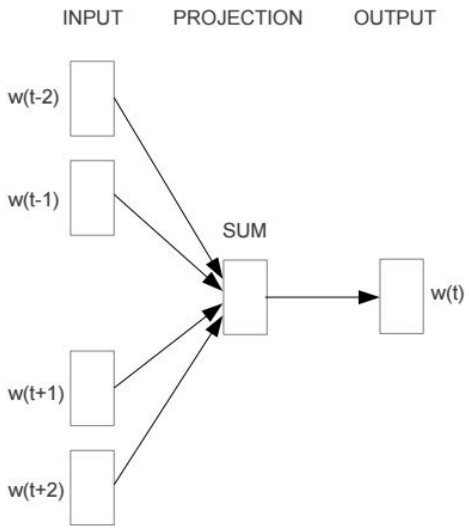
$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t \mid w_{t-n}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+n}).$$

Skip-gram - predicts surrounding words given the current word

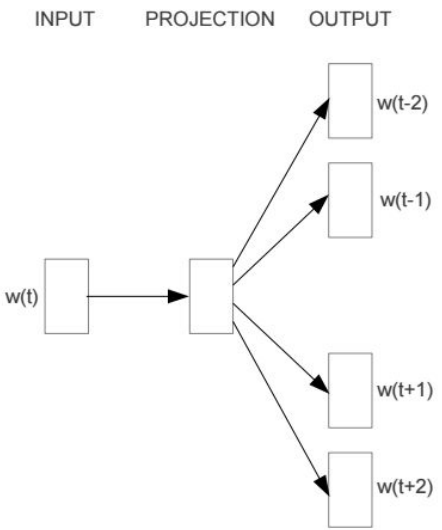
$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log p(w_{t+j} | w_t)$$

variables to optimize

denotes window range



CBOW

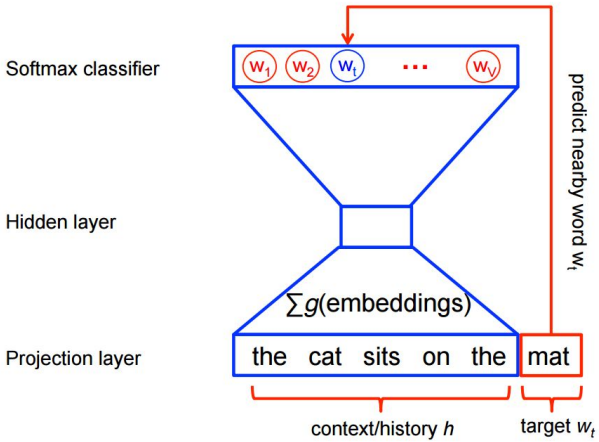


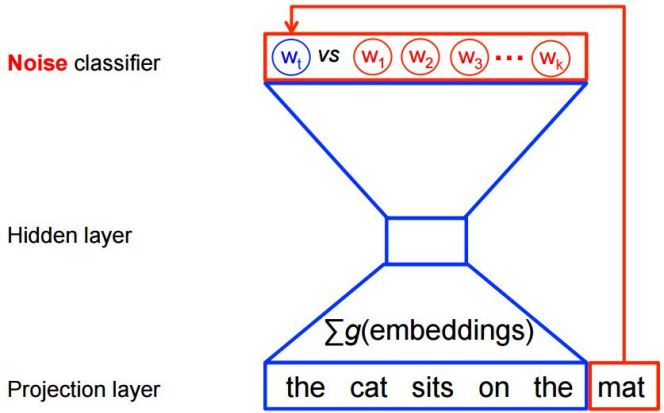
Skip-gram

Feedforward Neural Net Language Model (NNLM)

Paper source: https://arxiv.org/pdf/1301.3781.pdf

Quiz :)

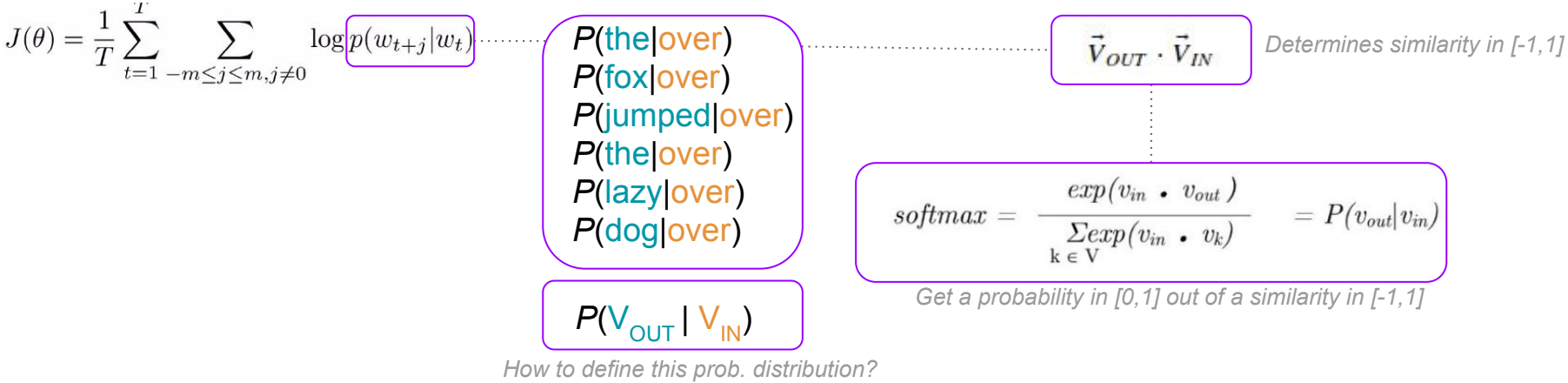

Softmax classifier
Hidden layer
Projection layer
w_1 w_2 w_t ... w_V
predict nearby word w_t
$\sum g$(embeddings)
the cat sits on the
mat
context/history h
target w_t


Noise classifier
Hidden layer
Projection layer
w_t vs w_1 w_2 w_3 ... w_k
$\sum g$(embeddings)
the cat sits on the
mat

Review Skip-gram architecture

Example: "The fox jumped **over** the lazy dog"

Objective function: maximize the likelihood of seeing the ***context*** words given the ***target*** word

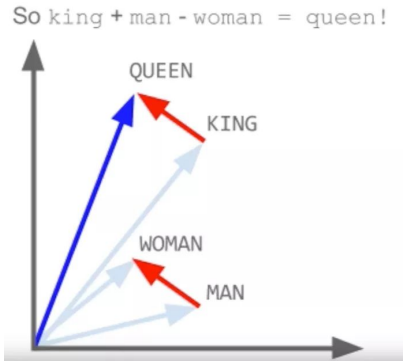


Hard work pays off

Features:

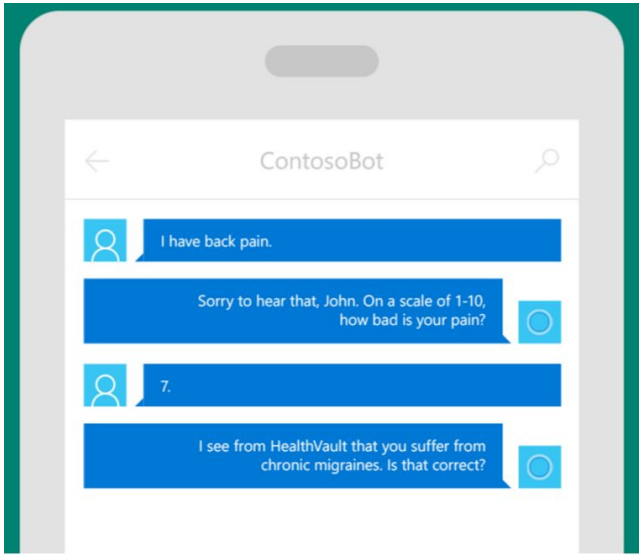
Vector Arithmetic.

```
In [77]: wv.most_similar(positive=['woman', 'king'], negative=['man'])

Out[77]: [('queen', 0.7118191719055176),
          ('monarch', 0.6189674139022827),
          ('princess', 0.5902431011199951),
          ('crown_prince', 0.5499460697174072),
          ('prince', 0.5377321243286133),
          ('kings', 0.5236844420433044),
          ('Queen_Consort', 0.5235945582389832),
          ('queens', 0.5181134343147278),
          ('sultan', 0.5098593235015869),
          ('monarchy', 0.5087411403656006)]

In [39]: X_test_word_average[:1].shape
```

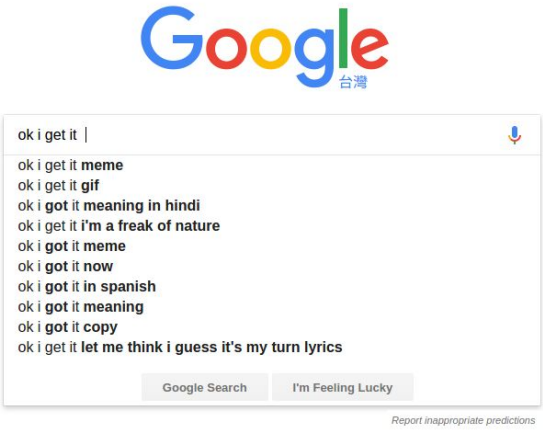


Application Opportunities

1. Smart Search engines
2. Context-aware conversational bots

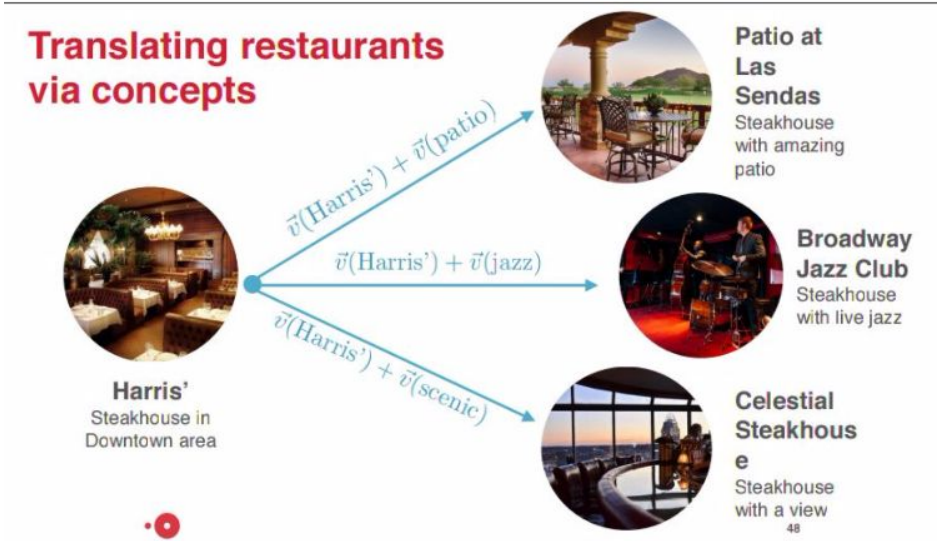


https://www.healthvault.com/en-us/health-bot/



Research Opportunities

1. Machine translation
2. Recommendation systems
3. Feature engineering



References

- Main Repository: https://goo.gl/ppHX65
- Other resources:
  - Gensim guide for word2vec: https://goo.gl/i2UrdH
- Original word2vec paper: https://goo.gl/7b72S9
- Stanford NLP with Deep Learning Course: http://web.stanford.edu/class/cs224n/syllabus.html
- Text Mining Overview: https://goo.gl/uNJDrs
- word2vec online calculator: http://rare-technologies.com/word2vec-tutorial/#app

Code Session

Sentence Classification

Task: Classify plots into one of six (6) movie categories

Data:

	Unnamed: 0	movieId	plot	tag
0	0	1	A little boy named Andy loves to be in his roo...	animation
1	1	2	When two kids find and play a magical board ga...	fantasy
2	2	3	Things don't seem to change much in Wabasha Co...	comedy
3	3	6	Hunters and their prey--Neil and his professio...	action
4	4	7	An ugly duckling having undergone a remarkable...	romance
5	5	9	Some terrorists kidnap the Vice President of t...	action
6	6	10	James Bond teams up with the lone survivor of ...	action
7	7	15	Morgan Adams and her slave, William Shaw, are ...	action
8	8	17	When Mr. Dashwood dies, he must leave the bulk...	romance
9	9	18	This movie features the collaborative director...	comedy

Data

In a future world devastated by disease, a convict is sent back in time to gather information about the man-made virus that wiped out most of the human population on the planet.

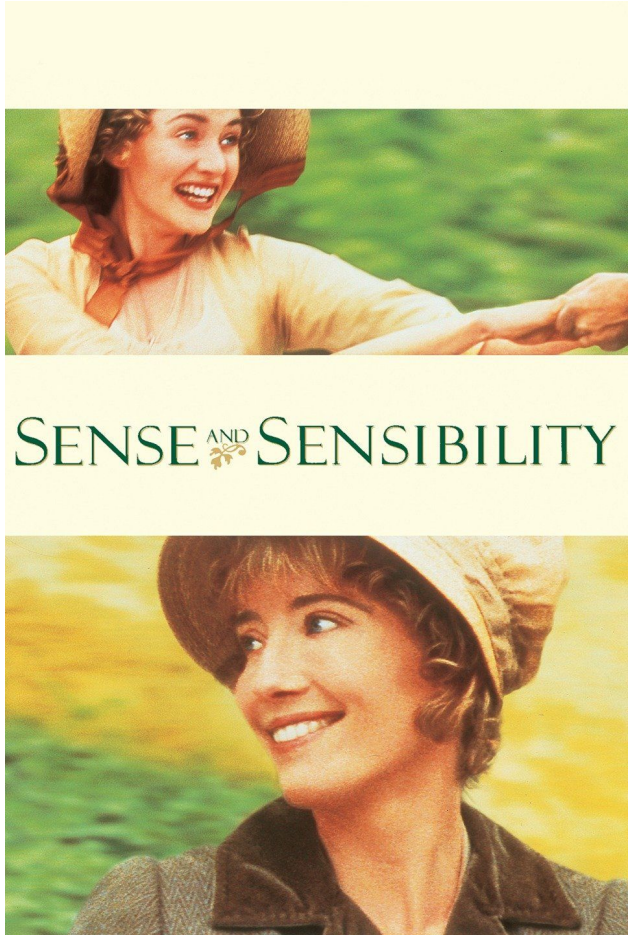
Sci-fi


BRUCE
WILLIS
MADELEINE
STOWE
BRAD
PITT
A TERRY GILLIAM FILM
12
MONKEYS
THE FUTURE IS HISTORY

Data

Mrs. Dashwood and her three daughters are left in straitened circumstances. When Elinor forms an attachment for the wealthy Edward Ferrars, his family disapproves and separates them. And though Mrs. Jennings tries to match the worthy (and rich) Colonel Brandon to her, Marianne finds the dashing and fiery John Willoughby more to her taste.

Romance



Project Ideas

1. **Emotion Analysis (a.k.a sentiment analysis)**
   - Data: https://goo.gl/KYacjz
   - Explore the data and provide visualizations
   - Apply text mining techniques learnt
   - Train a models to classify emotions
   - Provide evaluations
   - Prepare poster presentation
2. **Sentiment Analysis on Movie Reviews**
   - Data: https://goo.gl/JezgYg
3. **Annotation for Isaac**
   - Questionnaire: https://goo.gl/oCSgax

Training set:

for anger (updated Mar 8, 2017)
for fear (released Feb 17, 2017)
for joy (released Feb 15, 2017)
for sadness (released Feb 17, 2017)

Development set

Without intensity labels:
for anger (released Feb 24, 2017)
for fear (released Feb 24, 2017)
for joy (released Feb 24, 2017)
for sadness (released Feb 24, 2017)

With intensity labels:
for anger (released Apr 27, 2017)
for fear (released Apr 27, 2017)
for joy (released Apr 27, 2017)
for sadness (released Apr 27, 2017)

Note: For your competition submission for the test set, you are allowed

This is a *small* set of data that can be used to tune one's system, but
sure you try submitting your system output on the development set thr
that, well before evaluation period. Test data will have a format identica
Note: Since the dev set is small in size, results on the data may not be ir

Test set

Without intensity labels:
for anger (released May 1, 2017)
for fear (released May 1, 2017)
for joy (released May 1, 2017)
for sadness (released May 1, 2017)

With intensity labels:
for anger (released May 24, 2017)
for fear (released May 24, 2017)
for joy (released May 24, 2017)
for sadness (released May 24, 2017)

Future Projects

1. **Dashboard visualization to dynamically explore word embeddings**
   a. Build API: (Flask/Django recommended)
   b. Pretrained models: (Guide: https://goo.gl/5qt2Ki)
   c. Visualization: d3js / plotly / tensorboard
2. **Apply Deep Learning to text classification**
   a. LSTM - (Guide: http://colah.github.io/posts/2015-08-Understanding-LSTMs/)
   b. CNN - (Guide: https://goo.gl/PgLUs7)
   c. RNN - (Guide: https://goo.gl/5L9kci
3. **Vector Arithmetic Calculator**
   a. Starting point:https://rare-technologies.com/word2vec-tutorial#app