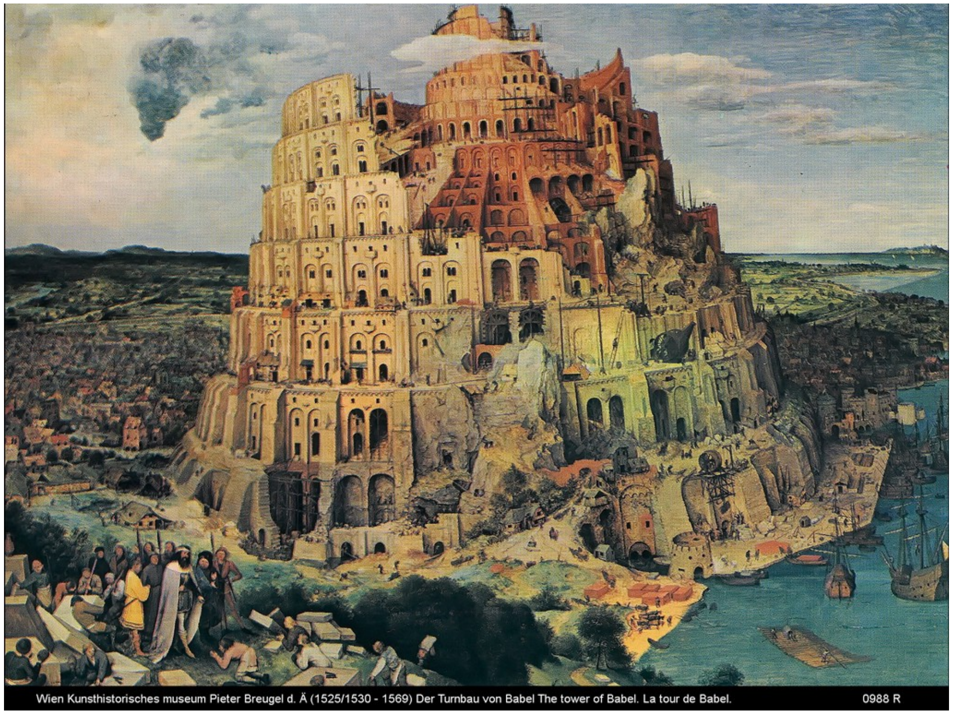
Teaching Advanced NLP to MSBA Students



Wien Kunsthistorisches museum Pieter Breugel d. Ä (1525/1530 - 1569) Der Turnbau von Babel The tower of Babel. La tour de Babel. 0988 R

NLP as a Rosetta stone for business insight in the digital tower of babel

Dokyun "DK" Lee

Assistant Professor of Business Analytics, Tepper School, CMU

DLforBusiness.com

I {Apply, Develop, Impact of} AI/ML For Business

Economics of Unstructured Data

Interpretable ML for Business

Unintended Consequence of AI

Courses Created From Scratch & Taught

- (Semester) Data Mining **Undergraduate** – R
- (Quarter) Data Mining **MBA** – R
- (Quarter) Machine Learning and NLP for Business Research **PHD Seminar** - Python
- (Quarter) Interpretable Machine Learning and Bias in Machine Learning **PHD Seminar**
- (Quarter) Deep Learning for Business: Mining Unstructured Data **MSBA** – Python

Courses Created From Scratch & Taught

- (Semester) Data Mining **Undergraduate** – R
- (Quarter) Data Mining **MBA** – R
- (Quarter) Machine Learning and NLP for Business Research **PHD Seminar** - Python
- (Quarter) Interpretable Machine Learning and Bias in Machine Learning **PHD Seminar**
- (Quarter) Deep Learning for Business: Mining Unstructured Data **MSBA** – Python

Deep Learning for Business: Mining Unstructured Data

- Teaching currently (Started May)
- MSBA students (our first cohort -domestic online only at this point 16 months – 16 quarter courses)
- **Format** - 7 weeks. 2 times per week (1 video lecture, 1 live vidyo conference lecture)
- Students Background
  - **Took python, machine learning, and deep learning in the curriculum before coming to my class**
  - All full-time working students

The Problem

- What to teach in 7-week period?
- Unstructured Data
  - Text, Images, Videos, Etc
- **Initially:** devised overview in all format
- **After a while**: realized it was impossible to make it good and cover all data type in 7 weeks.

Let’s Focus on Text. Why?

- Much business insight **still within text**
- First order info – In text, most of times
- Methods that work for text also applies to other form of sequence data
  - **Text**: sequence of atomic content
  - **Session/Clickstream**
  - **Purchase**: sequence of products
  - **Music**: sequence of notes
  - **etc**

Course Design Thoughts - Original Syllabus

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Key Points

- Understand the traditional & foundational NLP problems
- Make it easy to grasp the advantage of neural NLP & why we still need foundational NLP theory.
- The goal is business insight. Tools are useful iff students can understand how to use it to extract insight.
- Interpretable ML to make blackbox algorithms useful and make sense of unstructured data

Business Connection in Each Topic

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Motivation, value, and examples of business text data

Fundamental traditional NLP problems and tasks & connection to business

Ample amounts of current applications of NLP taken from

- **News**
- **Research Collaboration**
- **Capstone**
- **Previous experience as an ML engineer & NLP contractor**

Business Connection in Each Topic

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Often the first step to understand pattern in text data

- Unsupervised topic modeling
- **Supervised** version if y-var exists
- **Seeded LDA** to incorporate domain knowledge
- **Dynamic LDA** to see changes in topic over time
- **Structural Topic Modeling** to incorporate X-vars

- Students to run on their own fun data or business data

Business Connection in Each Topic

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Basics required to start neural net based nlp.

Embedding idea is applicable to any object – many business applications!

(More to come as I elaborate on this lecture as an example)

Business Connection in Each Topic

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Use tensorflow + keras to do many deep learning based models

Students already took deep learning.

My focus: what different models are good for and why they are useful for certain NLP tasks. What problems do certain deep models solve?

e.g., Non-linearities, Word order, Context, Memory

Business Connection in Each Topic

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Cover basic rnn generative model

Cover cutting edge models: Generative Pre-Training-2

Business example case: generated customized reviews, augmented intelligence (creative process).

Business Connection in Each Topic

1. Course Intro & Introduction to Text Mining
2. Traditional NLP
3. Latent Dirichlet Allocation: Topic Modeling
4. Word2vec & Doc2vec
5. Deep Learning Introduction & Tensorflow + Keras
6. Generative Model
7. Interpretable Machine Learning for Business Insight

Deep Learning Models= Blackbox

Need ways to poke & prod models

Need ways to explain the blackbox to obtain insight

Issues that arise from blackbox models

What could go wrong if ML is misused

Topics Covered – Original Syllabus

1. **Course Intro & Introduction to Text Mining**
   - Unstructured data value & motivations
   - What is text mining? NLP?
   - What is different now? Traditional NLP vs Neural NLP
2. **Traditional NLP**
   - Supervised learning
   - Subtasks (e.g., POS, NER, Parsing)
   - Basic Language Model & Naïve Bayes Classifier
3. **Latent Dirichlet Allocation: Topic Modeling**
   - LDA
   - Dynamic LDA, Correlated LDA, Seeded LDA, Supervised LDA, Structural Topic Modeling
4. **Word2vec & Doc2vec**
5. **Deep Learning Introduction & Tensorflow + Keras**
   - Multilayer Perceptron
   - Convolutional Neural Networks
   - Recurrent Neural Network
   - LSTM, GRU, LSTM-CNN
6. **Generative Models & Applications**
7. **Interpretable Machine Learning for Business Insight**

Course Work & Evaluation

- **Homeworks - 30%** – Total 2.
  - HW1 : Topic modeling, Word Embedding
  - HW2: Naïve Bayes (base) vs Deep Learning sentiment analysis
- **Timed Quizzes Online - 30%** – Total 5. 6-10Qs
- **Mini Group Work Presentation - 40%** - Total 3.
  - Replicate techniques learned on your own dataset & answer specific questions (topic model, generative model)
  - Last – open-ended project on your own data or company data (out-side of class consultation with me)

Course Format

- Video lecture (VL): technique details
- Reading Materials Given
- Live Vidyo session
  - **Bleeding-edge research**
  - **Current business applications** How are techniques used in business? Ample examples.
  - **Group work**: apply techniques & code covered on a novel dataset of student's own choosing
  - **Demo**: with code and jupyter notebook
  - **Discussion**: Method details, applying the methods, pitfalls of misusing the techniques

Specifics – **What Tools?**

1. **Python** – hands down there is no better language than Python to do NLP and ML in most cases (pandas, numpy, sklearn, etc throughout)
2. Traditional NLP – NLTK and spaCy package
3. Topic Modeling – Gensim + sklearn
4. Deep Learning Framework – Tensorflow + Keras
   - Alternatives pytorch (better for research and new models)
   - Mxnet (still growing), Caffe (mostly for image), etc
5. Interpretable ML – Various packages

Case in Point – W2V word embedding

1. Video Lecture: Introduce Concept and Math
   - **Behavioristic View** – Input: Corpus, Output: Words in Dense Vector, Preserve Sem Sim
   - **Conceptual Motivation** – distributional hypothesis
   - **Math & Computation & Nuances** – Skip-gram negative sampling likelihood function = mathematical formulation of distributional hypothesis. Compute through shallow net. Tantamount to co-occurrence matrix factorization with better tuning parameter.
   - **Extensions & Current State of The Art**– e.g., Doc2vec, *Space Embedding, BERT
2. Live Session
   - **Discussion** – business use of embedding – e.g., **Tin2VEC** and ASOS Customer Embedding Based on Clickstream logs
   - **Demo** – *Fun* Brand arithmetic based on GoogleNews embedding
   - **Research Focus** – Focused Concept Miner (interpretable deep learning based text method that combines topic model and embedding for business insight)
3. Homework – Students are given skeleton codes
   - Questions solving toy business problems

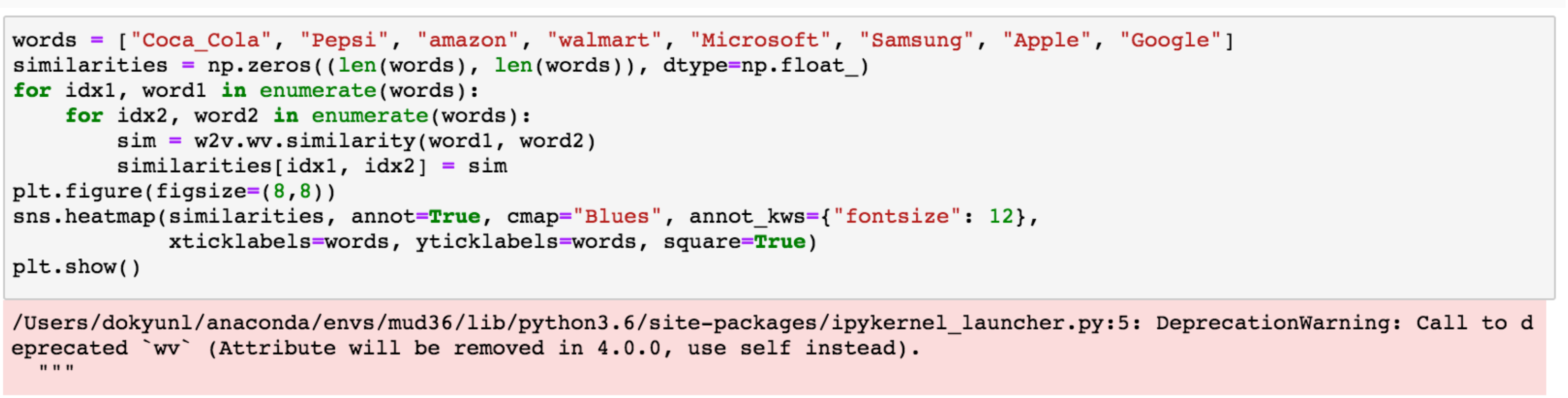
Case in Point – W2V word embedding

1. Video Lecture: Introduce Concept and Math
   - **Behavioristic View** – Input: Corpus, Output: Words in Dense Vector, Preserve Sem Sim
   - **Conceptual Motivation** – distributional hypothesis
   - **Math & Computation & Nuances** – Skip-gram negative sampling likelihood function = mathematical formulation of distributional hypothesis. Compute through shallow net. Tantamount to co-occurrence matrix factorization
   - **Extensions & Current State of The Art**– e.g., Doc2vec, Object2Vec, BERT
2. Live Session
   - **Discussion** – business use of embedding – e.g., **Tin2VEC** and asos Customer Embedding Based on Clickstream logs
   - **Demo** – *Fun* Brand arithmetic based on GoogleNews embedding
   - **Research Focus** – Focused Concept Miner (interpretable deep learning based text method that combines topic model and embedding for business insight)
3. Homework – Students are given skeleton codes
   - Questions solving toy business problems

Case in Point – W2V word embedding

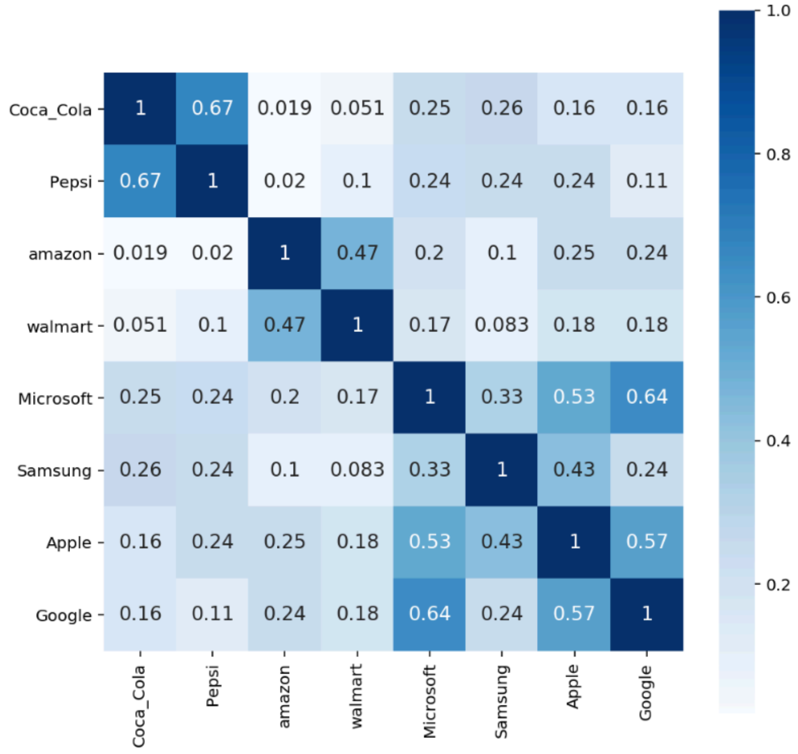
1. Video Lecture: Introduce Concept and Math
   - **Behavioristic View** – Input: Corpus, Output: Words in Dense Vector, Preserve Sem Sim
   - **Conceptual Motivation** – distributional hypothesis
   - **Math & Computation & Nuances** – Skip-gram negative sampling likelihood function = mathematical formulation of distributional hypothesis. Compute through shallow net. Tantamount to co-occurrence matrix factorization
   - **Extensions & Current State of The Art**– e.g., Doc2vec, Object2Vec, BERT
2. Live Session
   - **Discussion** – business use of embedding – e.g., **Tin2VEC** and ASOS Customer Embedding Based on Clickstream logs
   - **Demo** – *Fun* Brand arithmetic based on GoogleNews embedding
   - **Research Focus** – Focused Concept Miner (interpretable deep learning based text method that combines topic model and embedding for business insight)
3. Homework – Students are given skeleton codes
   - Questions solving toy business problems

Case in Point – W2V Demo



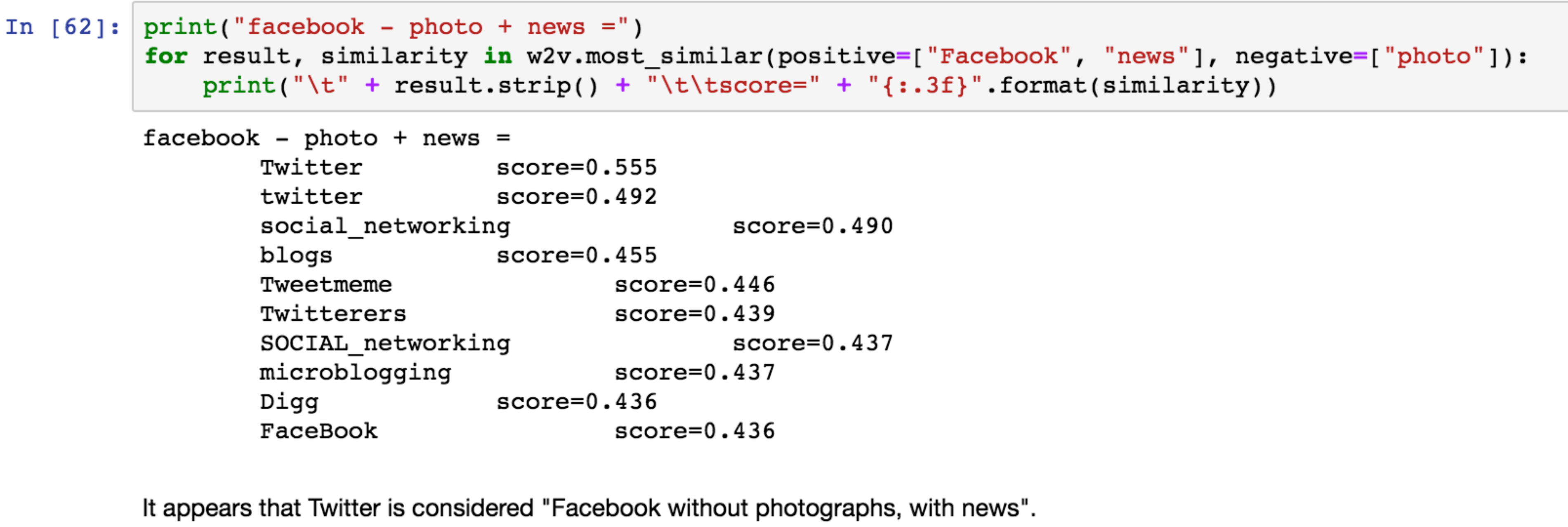
```
words = ["Coca_Cola", "Pepsi", "amazon", "walmart", "Microsoft", "Samsung", "Apple", "Google"]
similarities = np.zeros((len(words), len(words)), dtype=np.float_)
for idx1, word1 in enumerate(words):
    for idx2, word2 in enumerate(words):
        sim = w2v.wv.similarity(word1, word2)
        similarities[idx1, idx2] = sim
plt.figure(figsize=(8,8))
sns.heatmap(similarities, annot=True, cmap="Blues", annot_kws={"fontsize": 12},
            xticklabels=words, yticklabels=words, square=True)
plt.show()
```

```
/Users/dokyunl/anaconda/envs/mud36/lib/python3.6/site-packages/ipykernel_launcher.py:5: DeprecationWarning: Call to d
eprecated `wv` (Attribute will be removed in 4.0.0, use self instead).
  """
```



	Coca_Cola	Pepsi	amazon	walmart	Microsoft	Samsung	Apple	Google
Coca_Cola	1	0.67	0.019	0.051	0.25	0.26	0.16	0.16
Pepsi	0.67	1	0.02	0.1	0.24	0.24	0.24	0.11
amazon	0.019	0.02	1	0.47	0.2	0.1	0.25	0.24
walmart	0.051	0.1	0.47	1	0.17	0.083	0.18	0.18
Microsoft	0.25	0.24	0.2	0.17	1	0.33	0.53	0.64
Samsung	0.26	0.24	0.1	0.083	0.33	1	0.43	0.24
Apple	0.16	0.24	0.25	0.18	0.53	0.43	1	0.57
Google	0.16	0.11	0.24	0.18	0.64	0.24	0.57	1

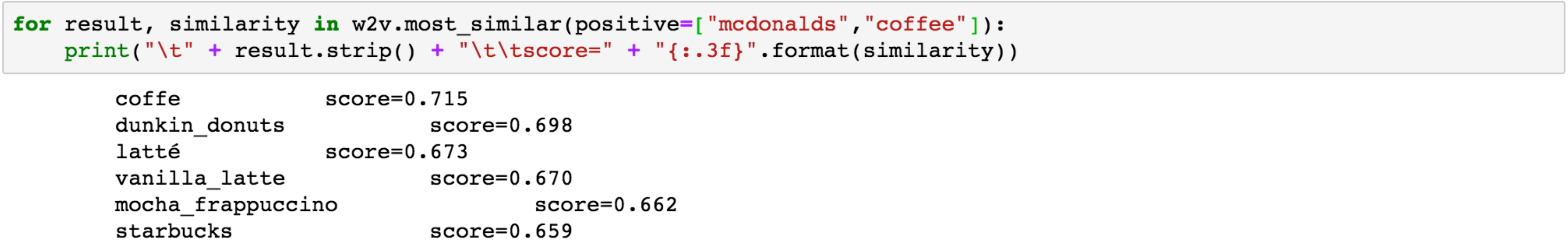
Case in Point – W2V Demo Continued



```
In [62]: print("facebook - photo + news =")
         for result, similarity in w2v.most_similar(positive=["Facebook", "news"], negative=["photo"]):
             print("\t" + result.strip() + "\t\tscore=" + "{:.3f}".format(similarity))
```

```
facebook - photo + news =
        Twitter         score=0.555
        twitter         score=0.492
        social_networking               score=0.490
        blogs           score=0.455
        Tweetmeme               score=0.446
        Twitterers              score=0.439
        SOCIAL_networking               score=0.437
        microblogging           score=0.437
        Digg            score=0.436
        FaceBook                score=0.436
```

It appears that Twitter is considered "Facebook without photographs, with news".

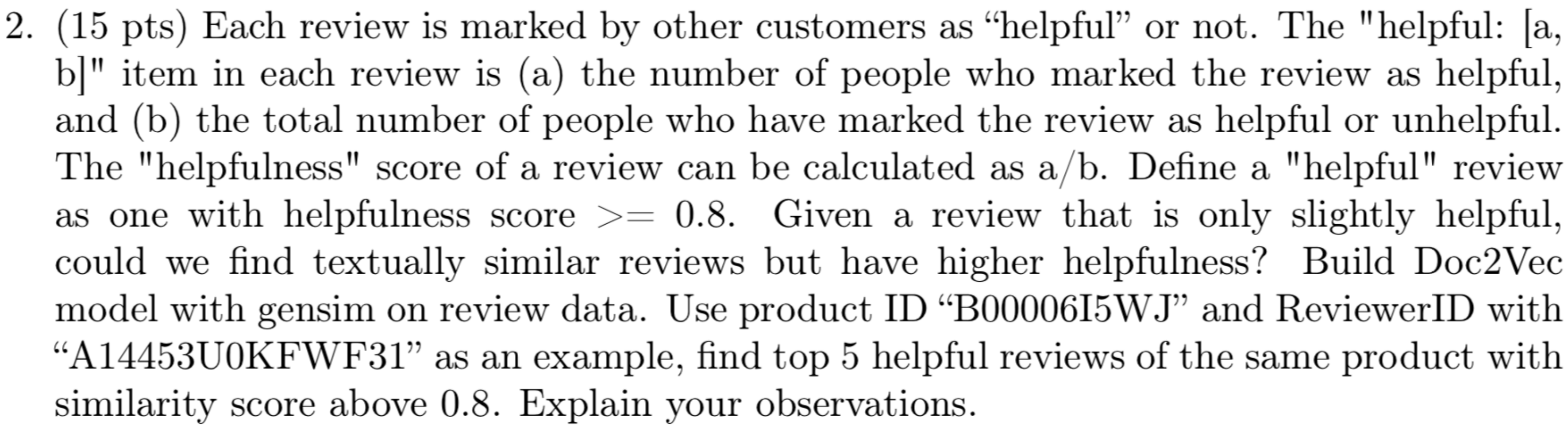


```
for result, similarity in w2v.most_similar(positive=["mcdonalds","coffee"]):
    print("\t" + result.strip() + "\t\tscore=" + "{:.3f}".format(similarity))
```

```
        coffe           score=0.715
        dunkin_donuts           score=0.698
        latté           score=0.673
        vanilla_latte           score=0.670
        mocha_frappuccino               score=0.662
        starbucks               score=0.659
```

W2V Homework – Amazon Review Data

Students are supposed to use techniques learned in class
To solve some toy version of business problems.
As shown below.



2. (15 pts) Each review is marked by other customers as “helpful” or not. The "helpful: [a, b]" item in each review is (a) the number of people who marked the review as helpful, and (b) the total number of people who have marked the review as helpful or unhelpful. The "helpfulness" score of a review can be calculated as a/b. Define a "helpful" review as one with helpfulness score >= 0.8. Given a review that is only slightly helpful, could we find textually similar reviews but have higher helpfulness? Build Doc2Vec model with gensim on review data. Use product ID “B00006I5WJ” and ReviewerID with “A14453U0KFWF31” as an example, find top 5 helpful reviews of the same product with similarity score above 0.8. Explain your observations.

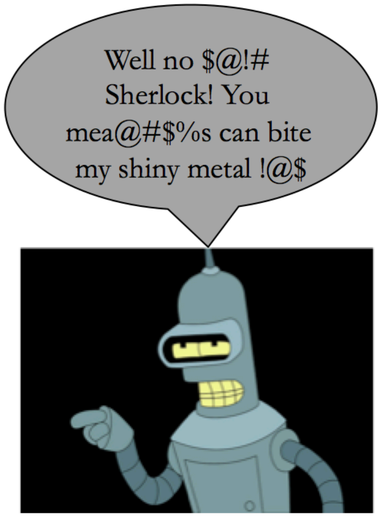
Bonuses Sprinkled Throughout – Fun Examples

Watson Gets An Attitude

IBM Watson learned urban dictionary in 2013…

"Watson couldn't distinguish between polite language and profanity -- which the Urban Dictionary is full of. Watson picked up some bad habits from reading Wikipedia as well. In tests it even used the word "bullshit" in an answer to a researcher's query.

Ultimately, Brown's 35-person team developed a filter to keep Watson from swearing and scraped the Urban Dictionary from its memory."

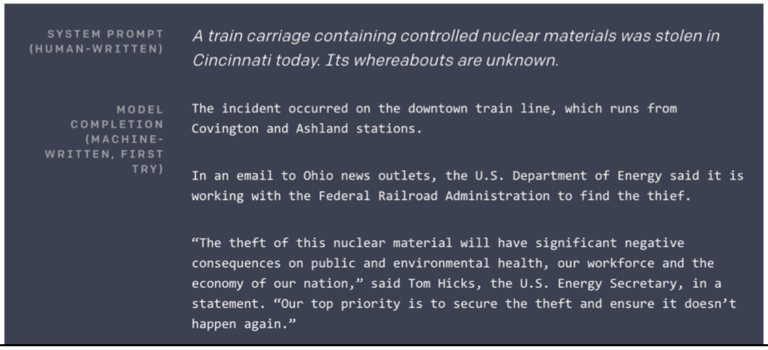

Well no $@!# Sherlock! You mea@#$%s can bite my shiny metal !@$

NLP is Hard! Example

- World leader in machine translation = Google.
- Still, this happens. Watch for laughs. https://www.youtube.com/watch?v=LMkJuDVJdTw

Fresh Prince: Google Translated - YouTube

Another Example: GPT-2 generative model



SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

```
The incident occurred on the downtown train line, which runs from
Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is
working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative
consequences on public and environmental health, our workforce and the
economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a
statement. "Our top priority is to secure the theft and ensure it doesn't
happen again."
```

Personalized Recommendations at tinder

The TinVec Approach

Steve Liu

Chief Scientist