Data Mining Lab

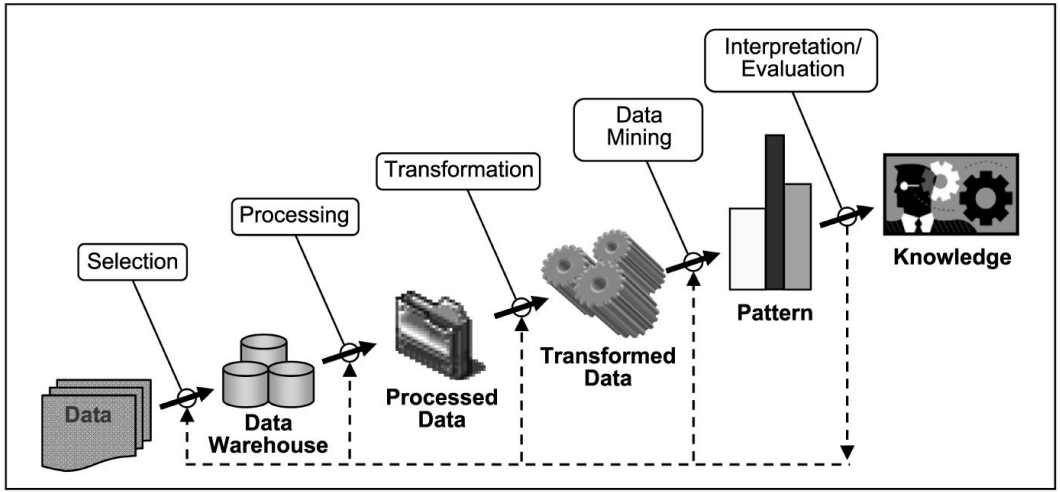
Fall 2017

Elvis Saravia
ellfae@gmail.com

king - man + woman = ?

Expectations for this lab

- Environment Setup
- Data preprocessing
- Training Models
- Evaluation of Models
- Assignment



Knowledge Discovery (KDD) Process

Word Vector Representations

Represent the meaning of a word?

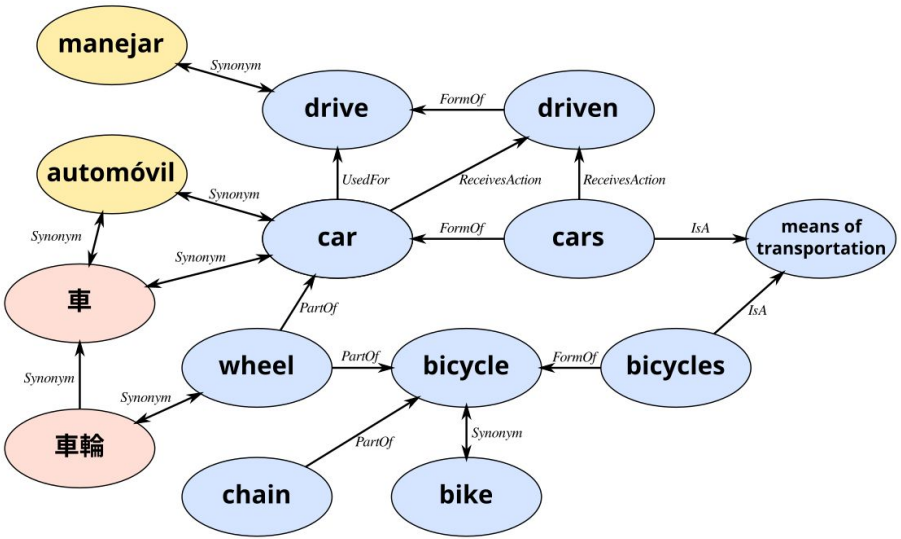
Words and phrases *directly represent* an idea

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?

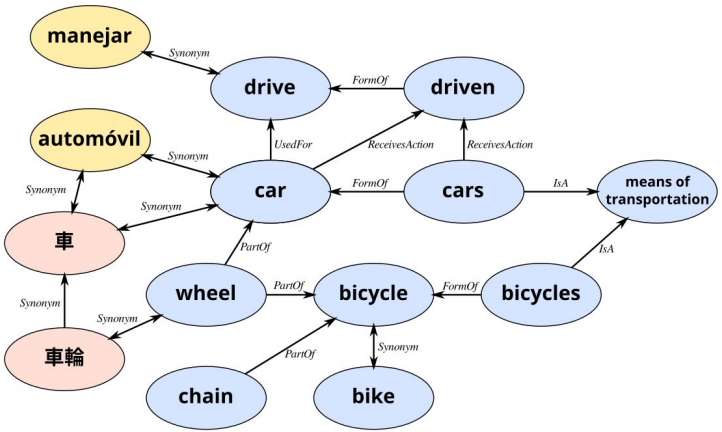
Solution: Taxonomy, such as WordNet and ConceptNet, that contains hypernyms (is-a) relationships and synonyms sets.



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]

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

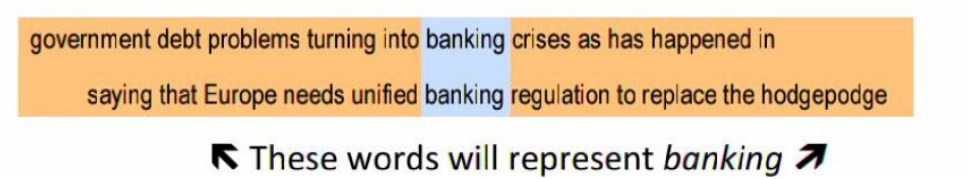
Distribution Similarity Based Representations

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)

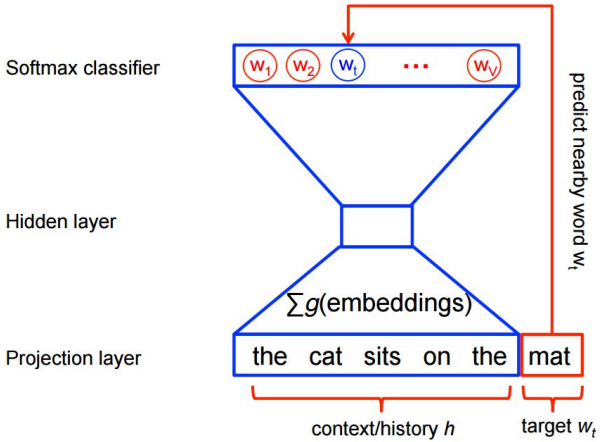
Ideas:

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

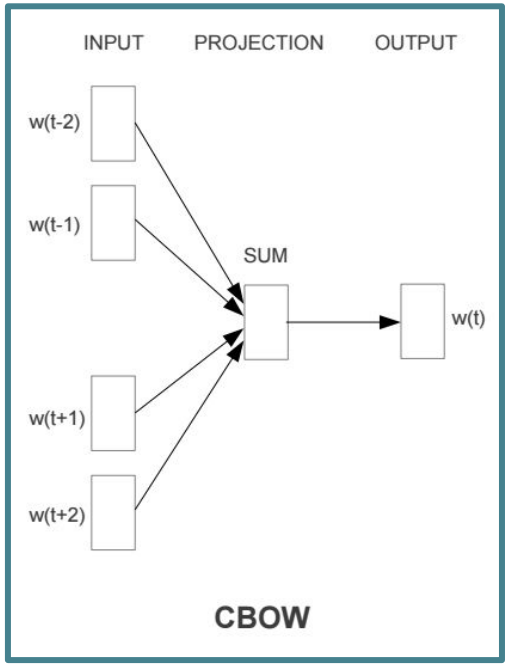
$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t \mid w_{t-n}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+n})$$

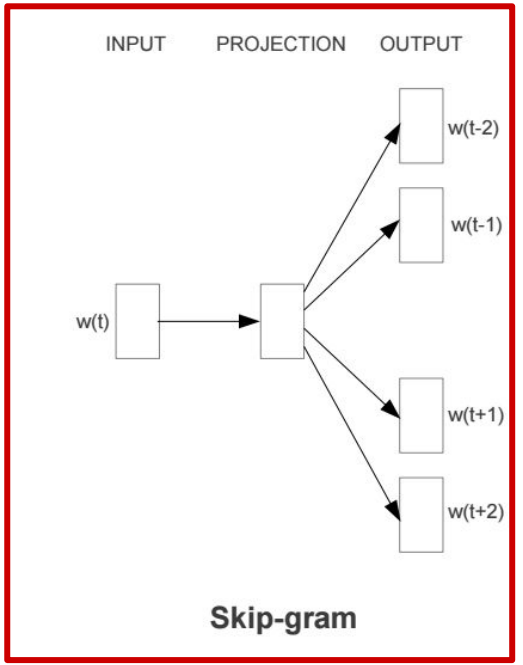
Skip-gram - predicts surrounding words given the current word

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

variables to optimize

denotes window range

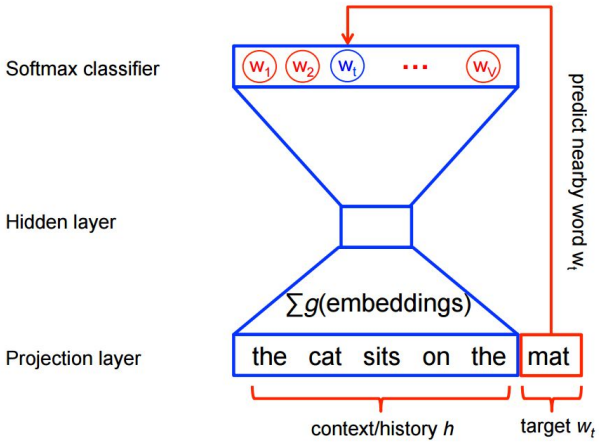


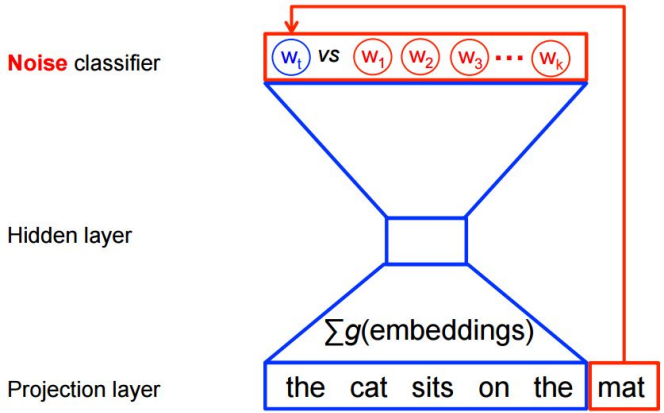


Feedforward Neural Net Language Model (NNLM)

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

Quiz :)

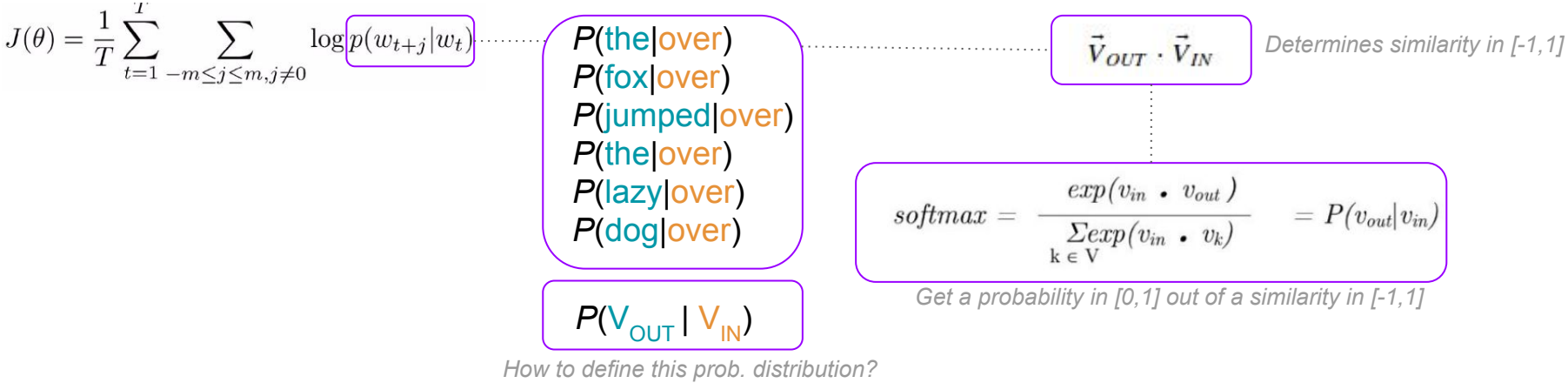

Softmax classifier
Hidden layer
Projection layer
w_1 w_2 w_t ... w_V
$\sum g$(embeddings)
the cat sits on the
mat
predict nearby word w_t
context/history h
target w_t


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

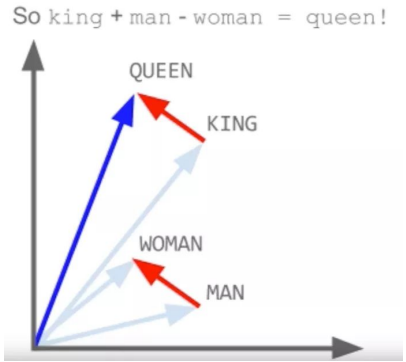


Hard work pays off

Features:

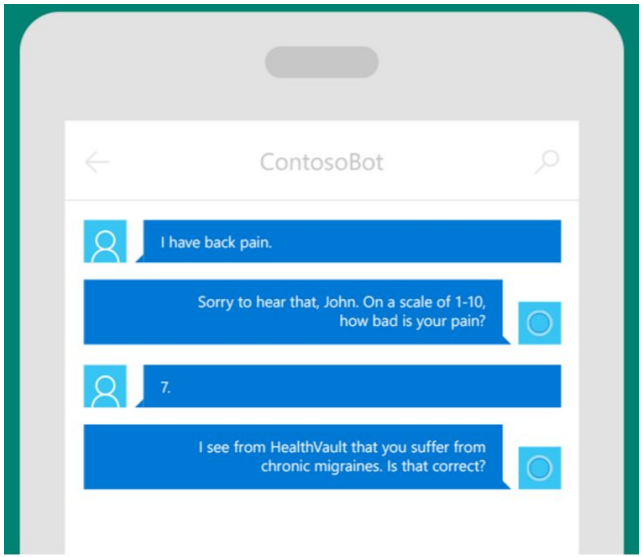
Vector Arithmetic.

```
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
```

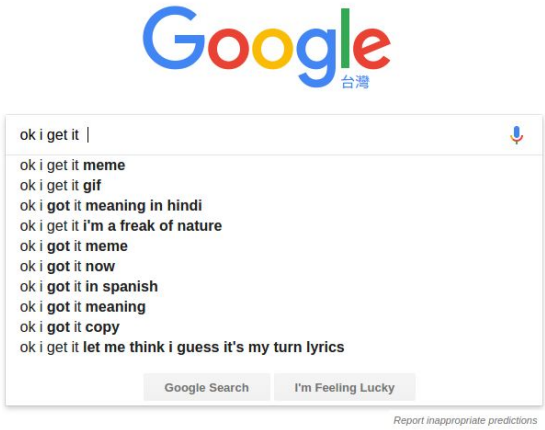


Application Opportunities

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

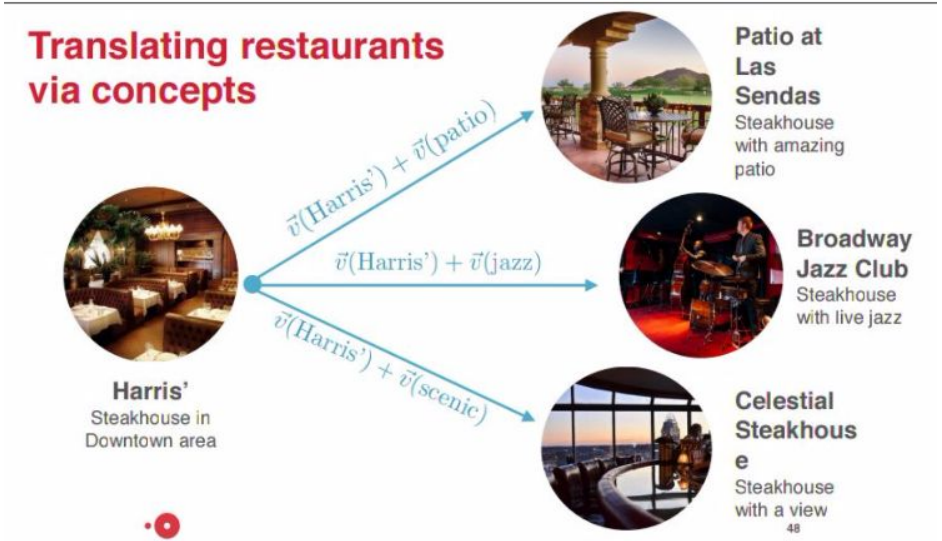


https://www.healthvault.com/en-us/health-bot/



Research Opportunities

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



References

- Main Repository: https://github.com/omarsar/data_mining_lab_fall_2
- 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 text into one of 4 emotions

Data: SemEval 2017 Task - Emotion Intensity

	id	text	emotion	intensity
617	20617	Recording some more #FNAF and had to FaceTime ...	fear	0.458
992	20992	@darwinwatersons @pennyfitzger31 @gumballwatte...	fear	0.271
144	20144	@Budget car rental you have made realize why ...	fear	0.729
224	20224	Retweeted Dr. Rand Paul (@RandPaul):\n\nStop f...	fear	0.667
385	40385	@SimonSSSJ123 @EllieG10853 @Onision @Eugenia_C...	sadness	0.485
574	10574	@MMASOCCERFAN @outmagazine No offense but the ...	anger	0.417
281	10281	Have wee pop socks on and they KEEP FALLING OF...	anger	0.562
579	30579	@Devilligan It's a beautifully sincere balanci...	joy	0.375
609	10609	I've been wanting salty fries from McDonald's ...	anger	0.396
231	30231	Ryan Gosling and Eva Mendes finally ; B joyfu...	joy	0.620

Data

You ever just find that the people around you really irritate you sometimes? That's me right now 😒

Anger 

You ever just find that the people around you really irritate you sometimes? That's me right now 😒

Data

r U scared to present in front of the class? severe anxiety... whats That r u sad sometimes?? go get ur depression checked out IMEDIATELY!!!

Fear 

r U scared to present in front of the class? severe anxiety... whats That r u sad sometimes?? go get ur depression checked out IMEDIATELY!!!

Demo