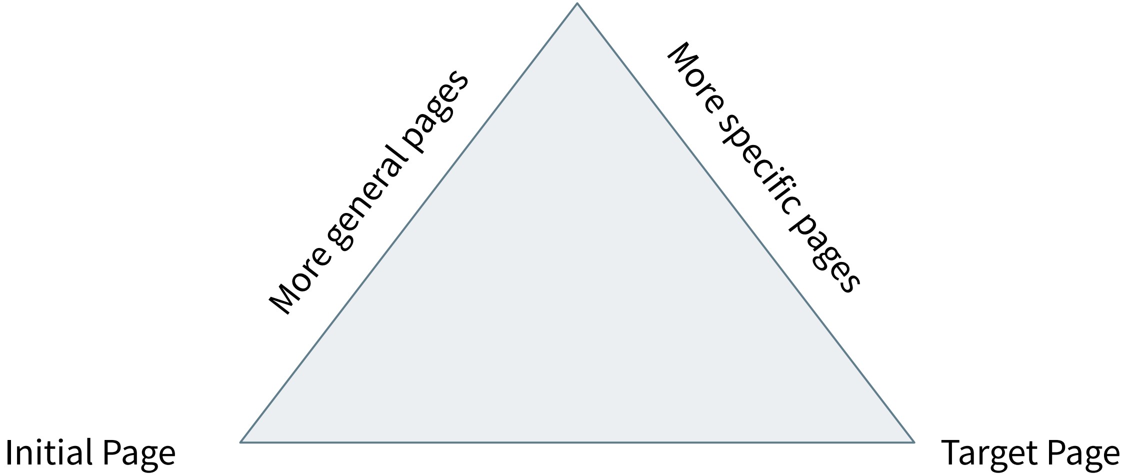
Playing the Wikipedia Game with Word Embedding

Ander Swartz

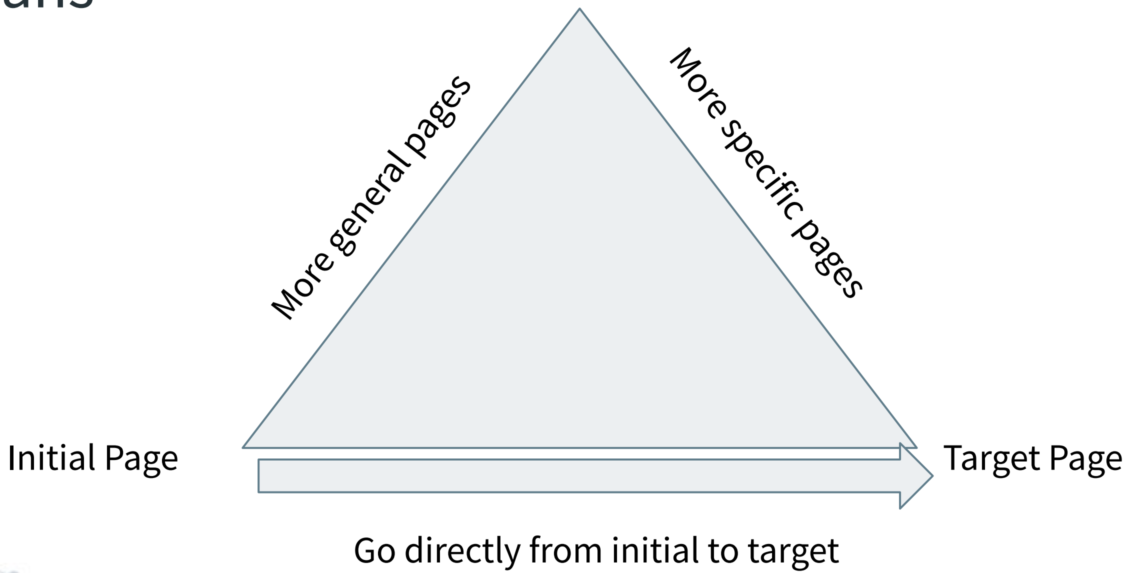
What is the Wikipedia Game?

- Get from one Wiki page to another in the least “clicks”, or shortest amount of time
- Usually played as a race with a friend



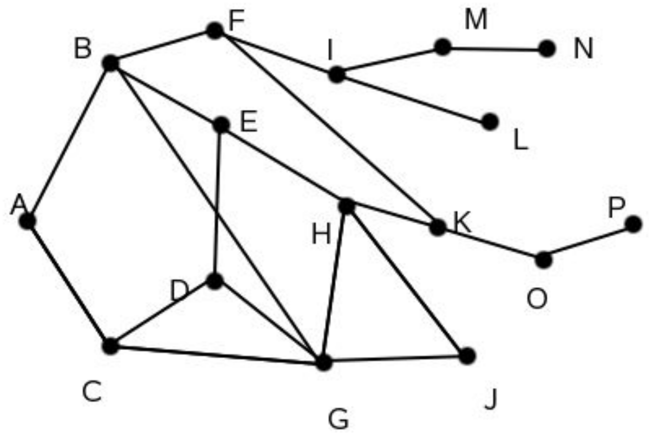
Can we do better with word embeddings?

- Goal: design an algorithm that is more direct than humans



Defining the Problem

- This can be solved by a search algorithm
- The state space is defined by all possible pages
- The “frontier” is all pages that can be reached from the current page
- How to select which page?
  - Heuristic function: Cosine similarity of word embeddings

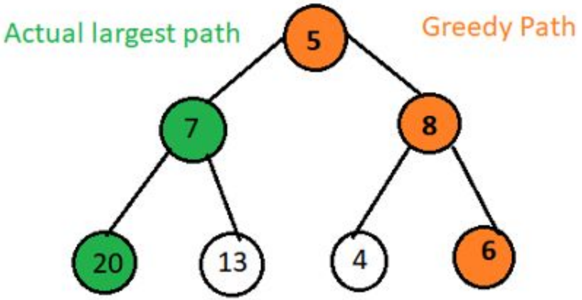


Which Word Embeddings to Use?

- I used Gensim's GloVe model
- GloVe uses a large matrix of word co-occurrence info
- It's task is to factorize the matrix to reduce the representation of each word as much as possible
- Unlike Word2Vec's CBOW or Skip-gram
- Both rely on distributional meaning
- My GloVe model was trained on a Wiki dump and the English Gigaword 5th Edition dataset

Back to the Search Algorithm

- Greedy Search: always choose state with highest score from heuristic
- Functions similar to depth-first search
- Is not always optimal!




```
Degrees of separation from Eukaryote --> Game : 20
['Life', 'Time', 'Season', 'History', 'Past', 'Minute', 'Second', 'Week', 'Da
y', 'Night', 'Summer', 'Baseball', 'Football', 'Nfl', 'Seattle', 'Dallas', 'Nh
l', 'Pittsburgh', 'Nba', 'Basketball']
```

Strengths and Weaknesses of Word Embeddings

Strengths

- Finding unique relationships between common words

Weaknesses

- Having less (or no) training on proper nouns

```
findDistanceSummingDisplay("Cat", "Computer")
```

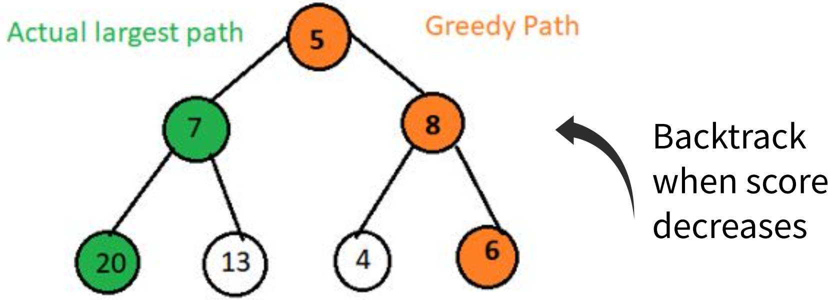
```
Initial Page: The cat (Felis catus) is a domestic species of small carnivorous
mammal.
Target Page: A computer is a digital electronic machine that can be programmed
to carry out sequences of arithmetic or logical operations (computation) automa
tically.
Current path: []
Pages visited: []
Microsoft 0.68780464
Current path: ['Microsoft']
Pages visited: ['Microsoft']
Computer hardware 0.9419043
Current path: ['Microsoft', 'Computer hardware']
Pages visited: ['Microsoft', 'Computer hardware']
Degrees of separation from Cat --> Computer : 2
['Microsoft', 'Computer hardware']
```

```
findDistanceSummingDisplay("Bay Furnace, Michigan", "African desert warbler'
```

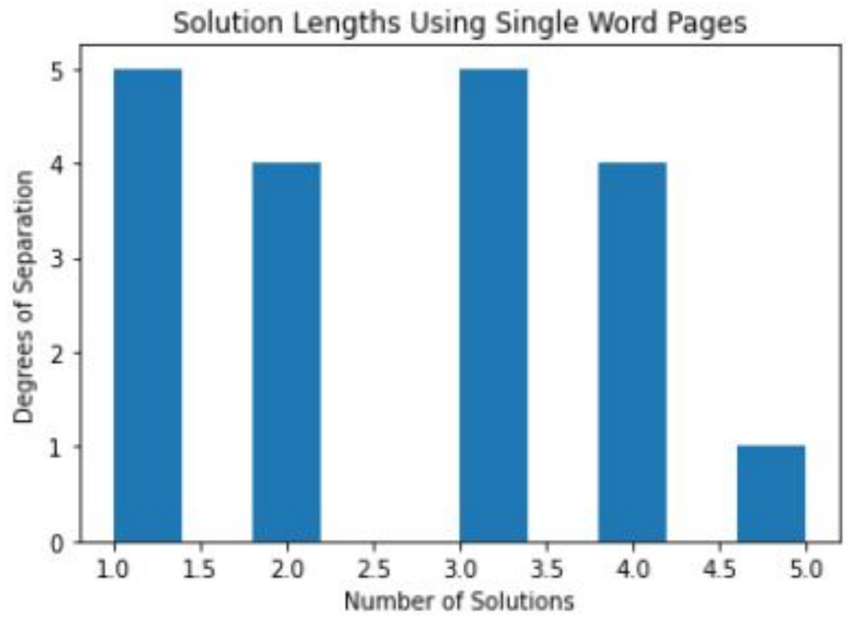
```
Pages visited: ['Isbn (identifier)', 'National library of south africa', 'Iz
iko south african museum', 'Southern africa mangroves', 'Temperate southern
africa', 'Karoo desert national botanical garden', 'South african institute
for aquatic biodiversity', 'Wildlife of south africa', 'African leopard', 'M
arine biodiversity of south africa', 'South african sendinggestig museum',
'South african english', 'White south african', 'White south africans', 'Sou
th african cuisine', 'South africa', 'South african airways', 'South african
navy', 'South african army', 'Southern afrotemperate forest']
List of Southern African indigenous trees and woody lianes 0.64338624
Current path: ['Isbn (identifier)', 'National library of south africa', 'Sou
thern afrotemperate forest', 'List of southern african indigenous trees and
woody lianes']
```

Modifications and Experiments

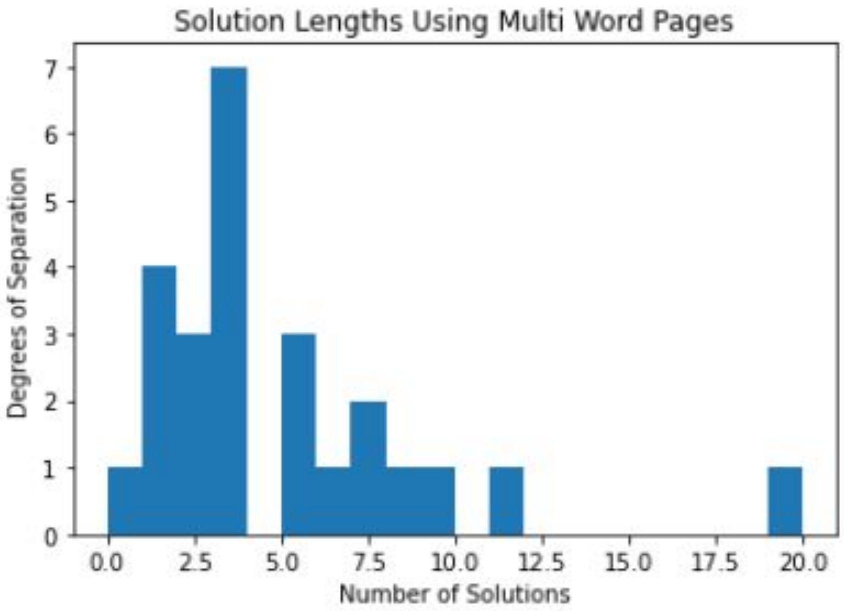
- Use multi-word pages
  - Word embeddings are for individual words
  - To handle a phrase, either average or add embeddings of words in a phrase
  - This is the main experiment
- Generalizing during beginning of search
- Backtracking



Results



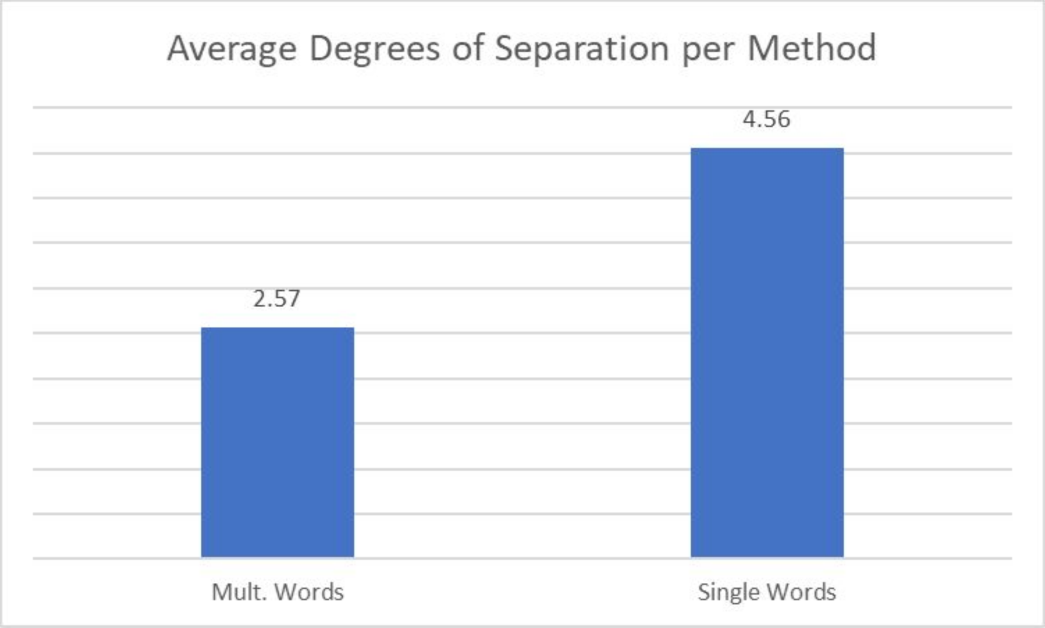
Solution Lengths Using Single Word Pages
Degrees of Separation
Number of Solutions
0
1
2
3
4
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0



Solution Lengths Using Multi Word Pages
Degrees of Separation
Number of Solutions
0
1
2
3
4
5
6
7
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Results

- Being able to use links with multiple words helped
- Drawbacks
  - Less time efficient: increases the search space
  - Will sometimes result in errors
  - In the Multi- Word column below, 6 entries were omitted because of errors



Some fun solutions

```
Degrees of separation from Empire --> Synthesizer : 2
['Cyrillic script', 'Qwerty keyboard']
```

```
Degrees of separation from Potato --> Stroke : 3
['Vitamin a', 'Ovarian cancer', 'Cancer']
```

```
Degrees of separation from Oxycodone --> Comedian : 1
['Stoner film']
```

```
Degrees of separation from Abstinence --> Chinatown : 4
['Counterculture of the 1960s', 'Williamsburg, brooklyn', 'Soho, manhattan', 'Chinatown, manhattan']
```

```
Degrees of separation from Cat --> Mat : 16
['Muhammad', 'Jonah', 'Adam', 'Khalifa', 'Malik', 'Quran', 'Ali', 'Wali', 'Turban', 'Helmet', 'Plastic', 'Rubber', 'Rattan', 'R
ope', 'Bed', 'Couch']
```

Conclusions

- Word embeddings have been demonstrated to be widely useful for semantic similarity
- Future work could explore using “backlinks” and bidirectional search
- Could also compare different word embedding models to see which work best