SJSU Students

PARAGRAPHS:

https://en.wikipedia.org/wiki/Killing_of_Harambe

On May 28, 2016, a three-year-old boy climbed into a gorilla enclosure at the Cincinnati Zoo and Botanical Garden and was grabbed and dragged by **Harambe**, a 17-year-old western lowland gorilla. Fearing for the boy's life, a zoo worker shot and killed Harambe. The incident was recorded on video and received broad international coverage and commentary, including controversy over the choice to use lethal force. A number of primatologists and conservationists wrote later that the zoo had no other choice under the circumstances, and that it highlighted the danger of zoo animals near humans and the need for better standards of care.

https://en.wikipedia.org/wiki/Donald_Trump

Born and raised in Queens, New York City, Trump attended Fordham University for two years and received a bachelor's degree in economics from the Wharton School of the University of Pennsylvania. He became the president of his father Fred Trump's real estate business in 1971, and renamed it to The Trump Organization. Trump expanded the company's operations to building and renovating skyscrapers, hotels, casinos, and golf courses. He later started various side ventures, mostly by licensing his name. Trump and his businesses have been involved in more than 4,000 state and federal legal actions, including six bankruptcies. He owned the Miss Universe brand of beauty pageants from 1996 to 2015. From 2003 to 2015 he co-produced and hosted the reality television series *The Apprentice*.

https://en.wikipedia.org/wiki/Bee_Movie

Bee Movie is a 2007 American computer-animated comedy film produced by DreamWorks Animation and distributed by Paramount Pictures. Directed by Simon J. Smith and Steve Hickner, the film stars the voices of Jerry Seinfeld, Renée Zellweger, Matthew Broderick, John Goodman, Patrick Warburton, and Chris Rock in supporting roles. The story follows Barry B. Benson, a honey bee who sues the human race for exploiting bees, after learning from his florist friend Vanessa Bloome that humans sell and consume honey.

https://en.wikipedia.org/wiki/Kevin_Bacon
Kevin Norwood Bacon[2] (born July 8, 1958)[3] is an American actor. His films include the musical-drama film Footloose (1984), the controversial historical conspiracy legal thriller JFK (1991), the legal drama A Few Good Men (1992), the historical docudrama Apollo 13 (1995), and the mystery drama Mystic River (2003). Bacon is also known for voicing the title character in Balto (1995), and was taking on darker roles, such as that of a sadistic guard in Sleepers (1996), and troubled former child abuser in The Woodsman (2004). He is further known for the hit comedies National Lampoon's Animal House (1978), Diner (1982), Tremors (1990) and Crazy, Stupid, Love (2011). His other well-known films are Friday the 13th (1980), Flatliners (1990), The River Wild (1994), Wild Things (1998), Stir of Echoes (1999), Hollow Man (2000), Frost/Nixon (2008), X-Men: First Class (2011), Black Mass (2015) and Patriots Day (2016). He is equally prolific on television, having starred in the Fox drama series The Following (2013–2015). For the HBO original film Taking Chance (2009), Bacon won a Golden Globe Award and a Screen Actors Guild Award, also receiving a Primetime Emmy Award nomination. More recently, Bacon portrayed the title character, and was the series lead, of the Amazon Prime web television series I Love Dick, for which he was nominated for a Golden Globe Award.

https://en.wikipedia.org/wiki/Shrek
Shrek is a 2001 American computer-animated comedy film loosely based on the 1990 fairy tale picture book of the same name by William Steig. Directed by Andrew Adamson and Vicky Jenson in their directorial debuts, it stars Mike Myers, Eddie Murphy, Cameron Diaz and John Lithgow as the voices of the lead characters. The film parodies other fairy tale adaptations, primarily aimed at animated Disney films.[6] In the story, an ogre called Shrek (Myers) finds his swamp overrun by fairy tale creatures who have been banished by the corrupt Lord Farquaad (Lithgow) aspiring to be king. Shrek makes a deal with Farquaad to regain control of his swamp in return for rescuing Princess Fiona (Diaz), whom Farquaad intends to marry. With the help of Donkey (Murphy), Shrek embarks on his quest but soon falls in love with the princess, who is hiding a secret that will change his life forever.

https://en.wikipedia.org/wiki/Amazon_rainforest
The Amazon rainforest, alternatively, the Amazon jungle[a] or Amazonia, is a moist broadleaf tropical rainforest in the Amazon biome that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 km2 (2,700,000 sq mi), of which 5,500,000 km2 (2,100,000 sq mi) are covered by the rainforest. This region includes

territory belonging to nine nations and 3,344 formally acknowledged indigenous territories.

https://en.wikipedia.org/wiki/Random-access_memory
Random-access memory (RAM /ræm/) is a form of computer memory that can be read and changed in any order, typically used to store working data and machine code.[1][2] A random-access memory device allows data items to be read or written in almost the same amount of time irrespective of the physical location of data inside the memory. In contrast, with other direct-access data storage media such as hard disks, CD-RWs, DVD-RWs and the older magnetic tapes and drum memory, the time required to read and write data items varies significantly depending on their physical locations on the recording medium, due to mechanical limitations such as media rotation speeds and arm movement.

https://en.wikipedia.org/wiki/Google
Google was founded in September 1998 by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University in California. Together they own about 14 percent of its shares and control 56 percent of the stockholder voting power through supervoting stock. They incorporated Google as a California privately held company on September 4, 1998. Google was then reincorporated in Delaware on October 22, 2002.[13] An initial public offering (IPO) took place on August 19, 2004, and Google moved to its headquarters in Mountain View, California, nicknamed the Googleplex. In August 2015, Google announced plans to reorganize its various interests as a conglomerate called Alphabet Inc. Google is Alphabet's leading subsidiary and will continue to be the umbrella company for Alphabet's Internet interests. Sundar Pichai was appointed CEO of Google, replacing Larry Page, who became the CEO of Alphabet. In 2021, the Alphabet Workers Union was founded, mainly composed of Google employees.[14]

https://en.wikipedia.org/wiki/DuckDuckGo
DuckDuckGo (also abbreviated as DDG) is an internet search engine that emphasizes protecting searchers' privacy and avoiding the filter bubble of personalized search results.[3] DuckDuckGo distinguishes itself from other search engines by not profiling its users and by showing all users the same search results for a given search term.[5] The company is based in Paoli, Pennsylvania, in Greater Philadelphia and has 124 employees as of January 2021.[2] The company name is a reference to the children's game duck, duck, goose.[6][7]

https://en.wikipedia.org/wiki/Zoom_Video_Communications

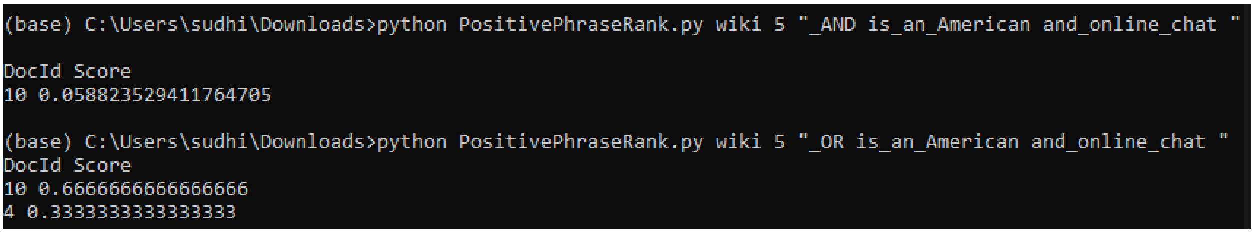
Zoom Video Communications, Inc. (or simply Zoom) is an American communications technology company headquartered in San Jose, California. It provides videotelephony and online chat services through a cloud-based peer-to-peer software platform and is used for teleconferencing, telecommuting, distance education, and social relations.[6][7] Eric Yuan, a former Cisco engineer and executive, founded Zoom in 2011, and launched its software in 2013.[8] Zoom's aggressive revenue growth, and perceived ease-of-use and reliability of its software, resulted in a $1 billion valuation in 2017, making it a "unicorn" company.[9] The company first became profitable in 2019,[10][11] and completed an initial public offering that year.[12] The company joined the NASDAQ-100 stock index on April 30, 2020.[13]

List of document topics:

1. Killing_of_Harambe
2. Donald_Trump
3. Bee_Movie
4. Kevin_Bacon
5. Shrek
6. Amazon_rainforest
7. Random-access_memory
8. Google
9. DuckDuckGo
10. Zoom_Video_Communications

Experiments 1-3: consisted of 2 terms for basic testing
Experiments 4+: consisted of 3 terms for HW intended output

Experiment #1:

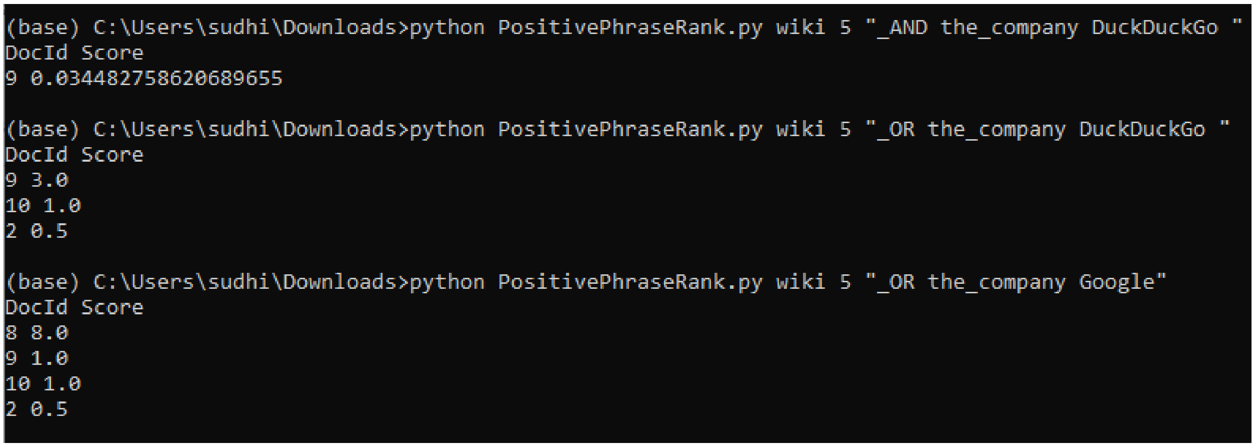


```
(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_AND is_an_American and_online_chat "

DocId Score
10 0.058823529411764705

(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_OR is_an_American and_online_chat "
DocId Score
10 0.6666666666666666
4 0.3333333333333333
```

We wanted to first test if our program can correctly run a purely conjunctive query or a purely disjunctive query within one document (Doc ID:10) versus another (Doc ID:4). The search is only amongst two terms as a base case but the search phrase is three words.

Precision@5: 1/1 ; 2/2
Recall@5: 1/1; 2/2

Experiment #2:

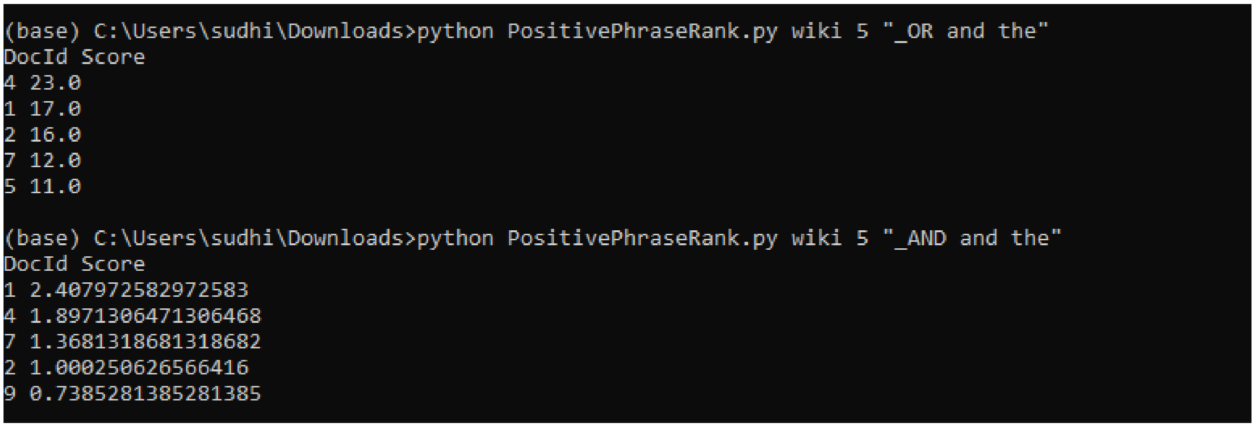


```
(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_AND the_company DuckDuckGo "
DocId Score
9 0.034482758620689655

(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_OR the_company DuckDuckGo "
DocId Score
9 3.0
10 1.0
2 0.5

(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_OR the_company Google"
DocId Score
8 8.0
9 1.0
10 1.0
2 0.5
```

For this experiment, the word “the company” involved at least 5 terms in the corpus. We formulated a search strategy using a conjunctive query and disjunctive queries tested the booleans between the documents. The outputs turned out as intended to when we wanted to exclude a phrase like “DuckDuckGo” or include “Google” with “the company”.

Precision@5: 1/1; 3/3; 4/4;
Recall@5: 1/1; 3/3; 4/4;

Experiment #3:

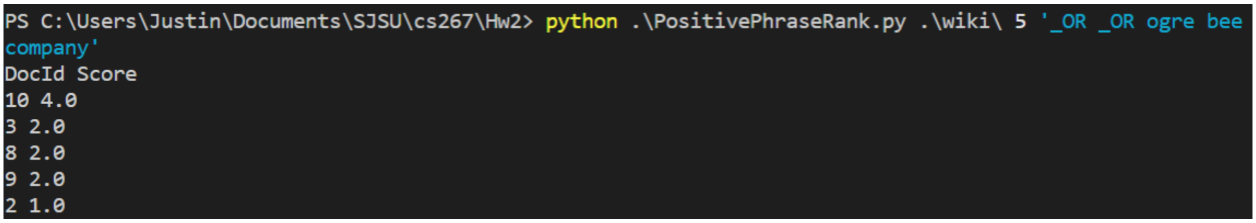


```
(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_OR and the"
DocId Score
4 23.0
1 17.0
2 16.0
7 12.0
5 11.0

(base) C:\Users\sudhi\Downloads>python PositivePhraseRank.py wiki 5 "_AND and the"
DocId Score
1 2.407972582972583
4 1.8971306471306468
7 1.3681318681318682
2 1.000250626566416
9 0.7385281385281385
```

For this experiment, we noticed “and” and “the” were the most frequent terms in the corpus for just document differentiation testing; we use a conjunctive query and a disjunctive query to observe the output differences between/with them in the query. Given that these terms are common prepositions and probably will be across all the documents, we expect the relevant results to be all documents with varying order when using boolean operators OR or AND.

Precision@5: 5/5; 5/5;
Recall@5: 5/10; 5/10;

Experiment #4 with 3 terms and a purely disjunctive query:



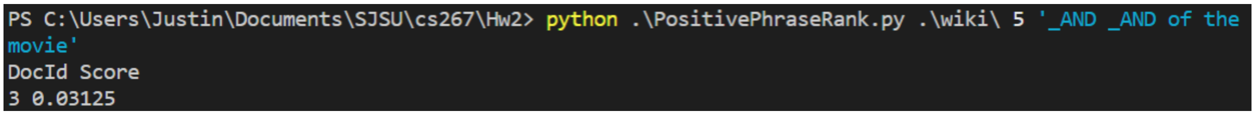
```
PS C:\Users\Justin\Documents\SJSU\cs267\Hw2> python .\PositivePhraseRank.py .\wiki\ 5 '_OR _OR ogre bee
company'
DocId Score
10 4.0
3 2.0
8 2.0
9 2.0
2 1.0
```

The query with these terms performed well. I expected the documents 2, 3, 5, 8, 9, 10 to be relevant to the query. Since the OR operand can only broaden our search results, it wasn't hard returning expected results. With only 5 scores shown I guessed Shrek might not make the cut. It indeed did not.

Precision@5 = 5/5

Recall@5 = 5/6.

Experiment #5 with 3 terms and a purely conjunctive query:



```
PS C:\Users\Justin\Documents\SJSU\cs267\Hw2> python .\PositivePhraseRank.py .\wiki\ 5 '_AND _AND of the
movie'
DocId Score
3 0.03125
```

The query with these terms performed below expectations. I expected the documents 3, 4, 5 to be relevant to the query. The AND query severely limits how many documents are returned since each paragraph needed to match all 3 terms. In order to return any document I need to insert stop words to match as many documents as I could. Of the documents returned only the 3rd document was assigned a score.

Precisoin@5 = 1/1

Recall@5 = 1/3

Conclusion:

Since we only had one paragraph from each wiki it was hard to find matches in general. There are few terms in one paragraph and even terms that I thought were relevant to a document didn't show up in the paragraph. There were also variations of terms in paragraphs that didn't match the terms in the query since the paragraph contained different variations of the root word. A stemmer was needed to find all occurrences of variations of a word and it's root. Disjunctive queries performed better since with small document sizes there aren't many terms. As long as there are no stop words you can return relevant results with greater accuracy. Conjunctive queries really limited the returned documents since the chances that all three terms are in one small paragraph are slim.