UIUC CS 512 Assignment 1

1. **top-k, medium-k and bottom-k phrases in single/multi-words rank list for each dataset, k = 30**

   (1) yelp single:

   - top:

   Moroccan, meatball, viet, swiss, yelper, armadillo, chimichanga, brewery, firehouse, castle, geisha, yelpers, stew, g's, disco, compass, Melrose, butte, si, Germany, cook's, lifetime, Illinois, vegas, cholla, hardy, firecracker, philly, village, farms

   - medium:

   Agony, laced, touches, rolling, heads, makers, raved, touched, walkable, complexity, menu.the, scalp, horrors, loan, drowning, survive, gastro, deliberation, resemblance, snuff, sickeningly, reek, holes, hope, brutal, jealous, nummy, capicola, raid, crossed

   - bottom

   Yourself, weren't, her, 09, containing, would, your, should, me, becomes, had, amongst, which, themselves, myself, ourselves, them, him, been, is, can, has, gives, itself, could, be, are, have, were, was

   (2) yelp multiple:

   - top:

   sake bombs, grand marnier, chocolate mousse, dairy queen, melting pot, chile relleno, papa johns, chilean sea bass, jobing.com arena, los betos, lobster tail, farmers market, palak paneer, pollo asado, al forno, ace hardware, carlsbad tavern, crab puffs, wild boar, kettle corn, filet mignon, cotton candy, pumpkin pie, la fitness, del rey, chop suey, cheesecake factory, pf changes, jamba juice

   - medium:

   a pear, the classic, the crispy, a nice restaurant, the value, this buffet, stuck to, past time, until 9, the decency, pointed to, no tip, table near, quite frequently, or safeway, the peanut sauce, you'll understand, the tempura, other froyo, s * * *, the pastry case, i'll drive, at richardson's, entry way, a special event, place to go for, hating on, bar side, love to, a small menu

- bottom

to wait long to, the delivery of, a cool spot to, ever eaten at, and just wanted to, and really wanted to, a darn good, of things to do, and very easy to, much food for, of people coming in, to dinner here, an afternoon of, good deals on, to lunch here, little place with, to drop by, to upgrade to, after checking in, in contact with, first heard about, a good reason to, only adds to, enough food to, a staple for, a stop at, while waiting on, to bite into, only issue with, little gem in

(3) nyt single:

- top:

nhl, sec, cia, iii, nba, gm, rbs, ipo, anc, gps, geneva, hsbc, dvd, hbo, adelaide, penn, columbia, redskins, llc, mlb, msnbc, sochi, wichita, isil, ncaa, Vienna, hiv, td, mvp

- medium:

Sympathetic, begging, mistreatment, stoppage, timing, modeled, consensus, slippers, microwave, productive, vending, suitcase, movement's, rationale, vocals, kirobo, brar, korelitz, wagstaffe, paperwhite, stab, lanning, rapfogel's, a.d.h.d, romário, demartino, kravchenko, rosenthal's, dot's, reyl

- bottom:

14.5, might, provides, must, 247, 151, 159, 304, 204, 146, is, are, can, which, would, who, be, whom, truly, 107, been, had, has, have, were, sincere, was, follows, ought, cannot

(4)nyt multiple:

- top:

freddie mac, carson palmer, jameis Winston, roy hibbert, thousand oaks, julie bataille, saratoga springs, dion waiters, berkshire Hathaway, marshawn lynch, cam newton, danny welbeck, beaver creek, janet yellen, kyrie irving, aston villa, valerie amos, gareth bale, wayne rooney, sidney crosby, shaun livingston, robin thicke, tiger woods, richie incognito, victoria azarenka, tesla motors, derek stepan, mariano rivera, kobe Bryant

- medium:

a turning point, an intern, on monday afternoon, onslaught of, strong enough, misstated part of, the Chinese, tens of millions of, a 21 point, thrives on, pauley said, of trayvon martin, equivalent of, don't think anyone, some chinese, a two run, four hearts, the nassau county, three categories, a preseason game, chairman and chief executive of,

well dressed, the guantanamo bay, the euro area, to insure, a tonne, air base in, 68 percent, the international institute. of jesus Christ

- bottom

a connection to, the interior of, no reason for, the agenda for, enough money to, an honor to, an employee at, the window of, still living in, a solution to, a picture with, an option for, the architect of, the gap to, a product of, the ceremony at, the condition of anonymity because of, after failing to, to speak about, a few hours after, still struggling to, all types of, 27 years in, another step in, not easy for, the pain from, first test in, to spend more, the question now, an account of

2. **A table includes number of qualified phrases in each dataset, average number of phrases in each sentence**

	Quality Phrases	**Num of Sentences**	**Avg hightlights**
nyt13_20k.txt	1506655	1593737	0.94536
YELP.100K.txt	1375808	1801910	0.763528

3. **Print several clusters and 20 words in each clusters. Describe Clustering method you are using and number centers you've set.**

- I use **K-means** to do cluster and set number centers as **100**. After doing word2vec, each word has a 100-vectore feature value. The worlds with smaller distance merge into same cluster.

(1) **nyt13_20k.txt**

- **Arts:** music, art, fashion, band, painting, rock, pop, dance, concert, opera, portrait, jazz, ballet, lyrics, piano, Dylan, country music, Lady Gaga, Madonna

- **Politics:** statement issued, press conference, Amnesty International, senior official, ambassadors, prime minister's, administration official, regional government, foreign affairs, White House spokesman Jay Carney, National Intelligence Service, Slovak, senior State Department, China's Foreign Ministry, internal affairs, National Bank, Chinese Foreign Ministry, regional leaders, Iran's foreign minister, executive secretary

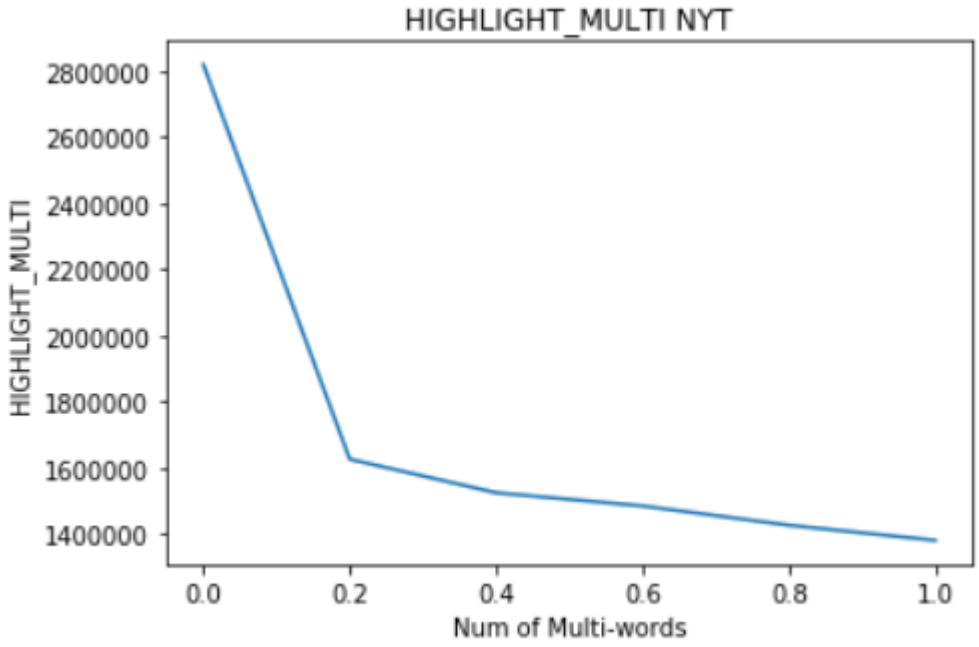
(2) **YELP.100K.txt**

- **Commerce:** Starbucks, Costco, Target, Whole Foods, Walmart, Trader Joe's, AMC, Circle, subway, Ikea, Nordstrom, Coach, Macy's, Walgreens, Dunkin Donuts, Lee Lee's, Sephora, Nike, H&M, costco

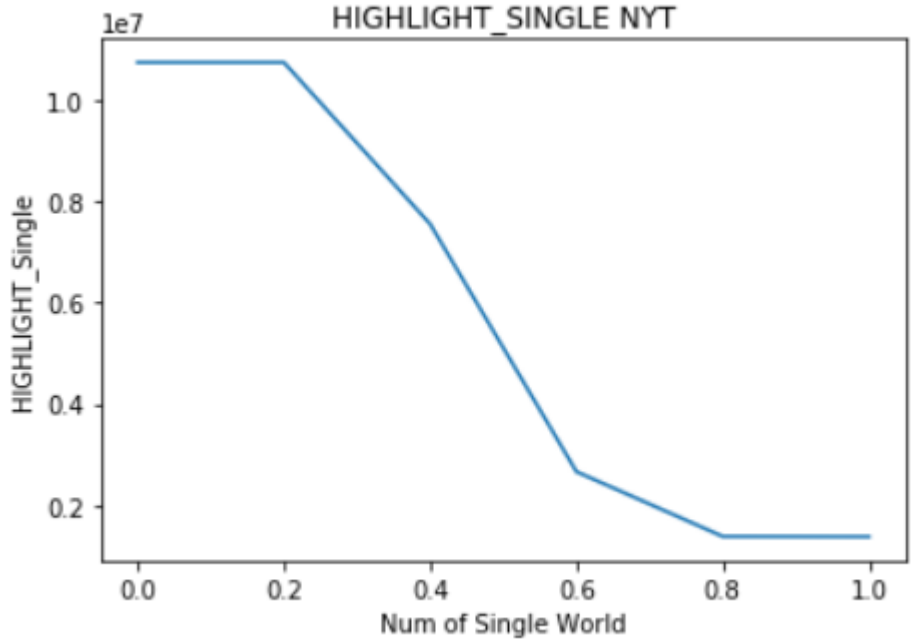
- **Food:** sushi, Mexican food, fast food, diner, Chinese food, cafeteria, comfort food, steakhouse, barbecue, soul food, home cooked, real food, junk food, sea food, Indian cuisine, fast casual, street food, traditional Irish, Italian cuisine, typical American

4. **A parameter study on AutoPhrase and Clustering, for example, by changing *HIGHLIGHT_MULTI,HIGHLIGHT_SINGLE* from** 0 to 1 in 0.2 increments **in phrasal_segmentation.sh you can get different number of phrases in corpus.** Draw a number of phrases versus ***HIGHLIGH_THRESHOLD*** curve for both multi-words and single word phrases returned by AutoPhrase**. By setting number of centers, you can get phrase clusters in different granularity.** Show some representative cluters and 10 words in each cluster for different granularity, e.g.(k=5, 10, 25)**.**

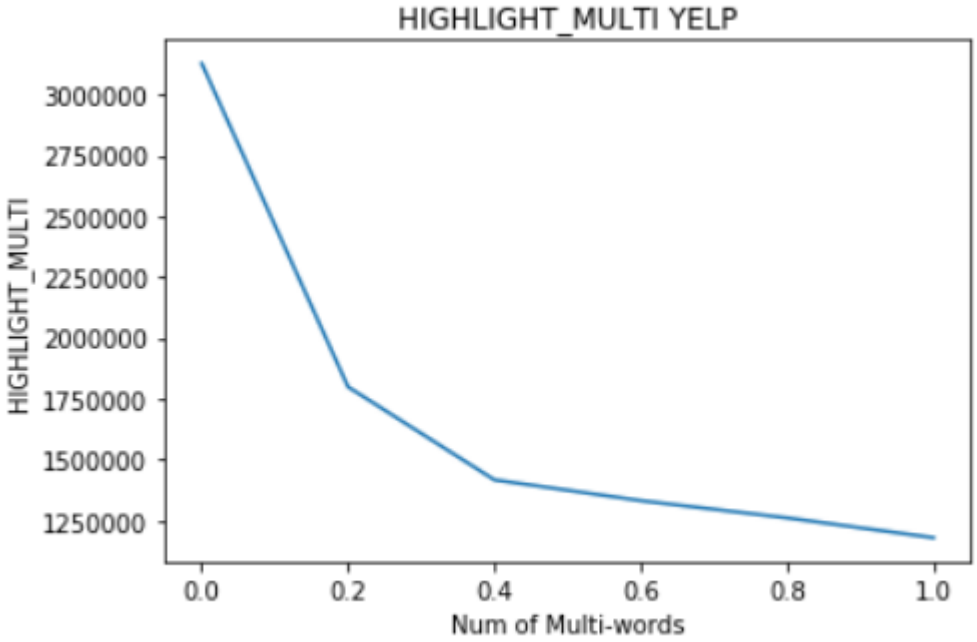
(1) NYT multi-words



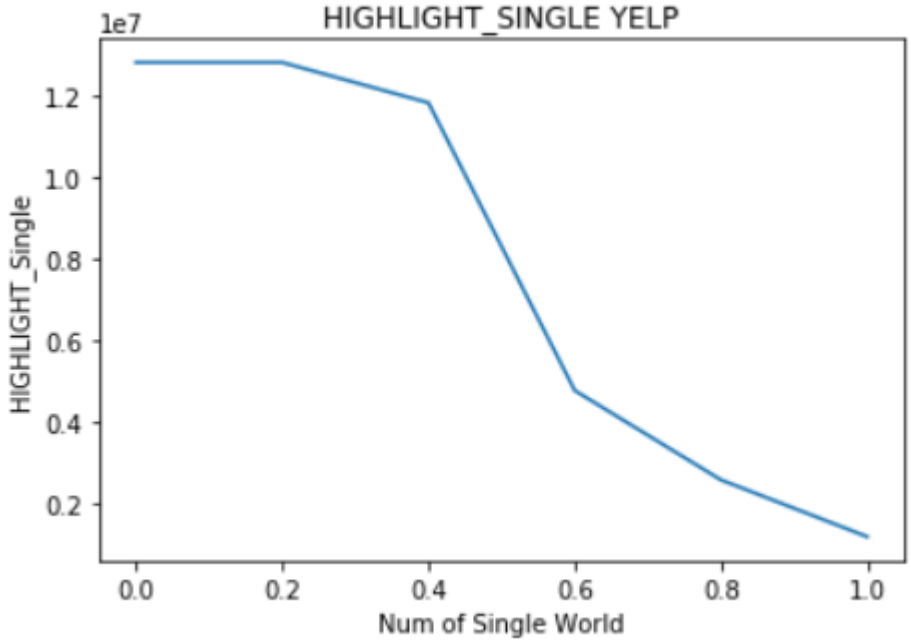
(2) NYT single word



(3) YELP multi-words



(4) YELP single word



When k is larger, granularity is also larger since words are more tend to different clusters.

(1) NYT

- **K=5:** York, president, Washington, director, Republican, school, center, House, Texas, lawyer
- **K=10:** police, family, report, court, news, federal, death, years ago, media, prison
- **K=25:** U.S, United States, country, American, China, Chinese, Iran, Russia, trade, Israel

(2) YELP

- **K=5:** food, love, order, friendly, pizza, bit, stars, taste, beer, coffee
- **K=10：** free, water, car, money, manager, dog, customer service, bartender, hair, talk
- **K=25:** live, music, game, watch, play, movie, sports, school, sound, games