CS512: Data Mining Principles, SP2018, HW 1

Nestor Alejandro Bermudez Sarmiento (nab6)

Introduction

The following report discusses the result of using AutoPhrase and Word2Vec over two different data sets to find quality phrases and later on group them using unsupervised clustering methods.

The data sets in question are a sample of the DBLP and YELP data sets commonly used in data mining performance reports.

The code used to perform the experiments can be found on GitHub [1]. Attached with this report you will find some of the results. These include the mined phrases (both single and multi words), the embeddings for the mined phrases and a file containing all the clusters and the corresponding phrases on each for each of the data sets.

Setup and Experiment Design

The experiments were performed in a macOS system; word2vec was downloaded and compiled using C; the clustering code was written in Python and the code to automatically run all the different setups was written in Bash. All the original code can be found in my GitHub repo.

The experiments consisted of trying different values for both *HIGHLIGHT_MULTI* and *HIGHLIGHT_SINGLE*. Due to the time it takes to run each setting only 25 settings were performed on each data set.
For both variables the values tried were: 0.1, 0.3, 0.5, 0.7, 0.9.

Results

This section includes some of the phrases found for each of the data sets, statistics about them and some examples of the clusters.

Lets first start by looking at some of the mined phrases for both single and multi word.

[1] https://github.com/nbermudezs/UIUC_CS512/tree/master/assignment_1

DBLP

Top-30 phrases (multiple words)

sun microsystems; disaster relief; authorship attribution; vickrey auctions; marching cubes; liner shipping; microsoft excel; spanning trees; turbo equalization; spreading activation; wind turbines; brazilian portuguese; bell labs; homomorphic encryption; google scholar; simultaneous multithreading; isabelle hol; noun phrase; multiset rewriting; liquid crystal; epipolar geometry; blood glucose; minimally invasive surgery; colon cancer; doctoral consortium; convex hulls; chip multiprocessors; ssl tls; sleep apnea; river basin;

Top-30 phrases (single word)

pearl; lai; lustre; ajax; inp; estelle; andrew; allen; mathematica; tracer; isis; mbps; vdd; powerpoint; yang; xen; finnish; stern; turkey; cisco; mason; jim; mosfets; lin; medline; hu; insar; taiwanese; promela; sparc;

Middle-30 phrases (multiple words)

binary moment; von mr; fundamental aspect; synchronous data; an important building block; simulation of; single data set; explicit word; obey certain; 2005 2006; at different sites; randomized distributed; zero characteristic marriage between; systems for strategic; scalability and dynamic; systems for digital; networking and applications; applications on embedded; applications and platforms; dynamics in complex; systems for industrial; applications of geometric; applications for small; stability for systems; computing in java; systems for effective; control and routing; developed and combined; techniques in reducing;

Middle-30 phrases (single word)

preferences; influencing; jscc; fre; unobserved; v2.0; kommerzielle; murmurs; topaz; imprints; clavier; troubleshooter; pseudotriangulations; dramaturgical; separatrix; prospection; bubbly; tree's; spiculated; misfolding; conciliating; lizard; montages; percolating; condensate; reputational; elementwise; illocutionary; hunger; muscl;

Bottom-30 phrases (multiple words)

and effective approach to; the theoretical limit of; a standard database of; the typical characteristics of; a specific feature of; an optimal selection of; the asymptotic capacity of; the basic mechanisms of; a common interface to; a fading channel with; a program logic for; a prior distribution on; a fundamental change in; this algorithm runs in; the creation and manipulation of; a requirements analysis for; a necessity to; an implementation model of; the knowledge stored in; to fault tolerance in; the inherent uncertainty in; for relevance feedback in; the data gathered by; a uniform approach for; the underlying geometry of; an estimation method of; the total number of processors in; the factors leading to; the concepts presented in; the topics discussed in;

Bottom-30 phrases (single word)

Îšo; cktsteiner; fotw; votc; Îż3; pieceware; udupa; lpoll; iszero; changelog; mkj; blogreg; pcvms; glazier; syncsql; rpprimary; pp2a; ifml; geeve; qlab; wgmww; modc; bcri; cvwann; pidis; arwon; sadr; laac; pcsw; rfpf;

YELP

Top-30 phrases (multiple words)

noble beast; tom's thumb; papago park; farmers market; fox news; jason's deli; fettuccine alfredo; golden corral; xtreme bean; eggs benedict; britney spears; pub crawl; loco patron; la canasta; mason jar; anchor steam; fox concepts; bon appetite; gainey village; palm springs; ajo al's; tammy coe; pollo fundido; daily dose; brussels sprouts; hobby lobby; barrio queen; coney island; frank sinatra; blue moon;

Top-30 phrases (single word)

redbox; coach; ale; mormon; samurai; hef; serrano; raspberry; police; antipasto; yelper; naan; lululemon; smith; northern; chimichanga; mimosa; sochu; latin; ristorante; mocha; sichuan; meatloaf; philly; magazine; sirloin; fridays; fondue; panini; mole;

Middle-30 phrases (multiple words)

the feta; a bad seat; not so nice; 2 x; for mother's day; more than one star; a quick serve; a drug store; planned on getting; this aj's; a video game; haven't already; a bit doughy; bf liked; you'll ever taste; a much needed; yelp about; very chill; increase in; at bobby q; mind paying more for; sucker for; a bit snooty; a topping; we'll definitely go back; no option; to solve; 3 to 4 stars; a loooong; a trailer;

Middle-30 phrases (single word)

td; ots; aimlessly; growlers; za'atar; collector; milder; committed; backside; resulting; granita; frequency; indulging; panel; grrrrr; distant; sprung; yasha; schwag; tracker; ravs; locating; mahogany; puzzled; customer; raffle; lps; fellas; cubes; scrambled;

Bottom-30 phrases (multiple words)

a great place to come with; even longer for; want to stop in; any place i've; get busy but; five minutes to get; to bring back to; and it's easy to; to play with and; a little small and; little patio with; so long to go; so easy to get; a little extra to; the staff does not; to agree to; to head back and; to feel like you're; not always easy to; to live here; so cool to; the manager came by to; it's amazing to; really really wanted to; take long to get; a perfectly made; no reason to go to; any other place in; this location for about; to go inside to;

Bottom-30 phrases (single word)

were; be; was; is; are; could; has; have; been; itself; can; ourselves; which; gives; your; had; would; should; ought; him; causes; merely; containing; werent; am; them; arent; |; might; 07;

Statistics

For each of the settings of the experiment, the number of phrases segmented and the average number of phrases on each sentence was capture. The results in the following table show the metrics for the recommended setting (HIGHLIGHT_MULTI = **0.5**, HIGHLIGHT_SINGLE=**0.9**).

	Data set	
	DBLP	YELP
# Phrases	7,877,383	1,587,940
Avg. Phrases/Sentence	0.597369	0.38297

Table 1: Number of phrases and average per sentence.

Clusters

Although multiple clustering methods exist, DBSCAN, k-Means, AGNES, I decided to use spherical k-Means, which is nothing more than the regular k-Means using **cosine** as the distance measure. Experiments were performed using 10, 20, 40 and 80 cluster centers. For the sake of pragmatism only the phrases selected from the clustering with 40 centers are shown in this report but everything else can be found in the GitHub, including best effort labels for some of the clusters. Some interesting finds will be discussed in this section.

DBLP

Cluster 29 (Databases/Data Mining):

POS Tagging	Information Extraction	Sequence Database
Frequent Pattern Mining	Feature Mining	Semi-Structured Data
Text Corpora	Spatio-Temporal Data	Image Retrieval Systems
Heterogeneous Data	Graph Structure	Database Views
Sequential Pattern Mining	Biological Networks	Mining Algorithms
Knowledge Extraction	Database Theory	Bitemporal Databases
Data Preparation	Data Clustering	

Cluster 15 (German words):

QualitÃď't der	Ein objektorientierter	Neue Herausforderungen
KÃijnstliche Intelligenz	Die Programmiersprache	Zur Erfassung
Diagramas de	Bedeutung von	Netze mit
Techniken und	Und Semantik	Kosten und
Zur Theorie der	Eine Herausforderung	Nach der
AnsÃďtze fÃijr	Des Wissensmanagements	Erste Erfahrungen
Modell und	Integrierter Ansatz	

Cluster 17 (Networks):

Wi-Fi	Real-time traffic	QoS guarantees
TCP/IP	QoS provisioning	TCP-based
Ad hoc networks	Multicast protocols	WiMAX-networks
Packet level	Local area networks	Optimal routing
Low bandwidth	IP traffic	WPANs
Congestion control	Packet routing	3G networks
ATM networks	Distributed Hash Tables	

Cluster 23 (Embedded Systems):

FPGA-based	Low latency	File access
Multi-core	GPU-based	Bus-based
Embedded system	AMD	Dual-core
Floating-point	Cache memory	High Performance Computing

Hardware Architectures	Multi-core architectures	Per cycle
Read-write	FPGA-technology	Block transfer
NoC-based	NAND flash	

Cluster 34 (Artificial Intelligence):

Support vector machines	Random Forest	Naive Bayes
Binary classification	Graphical model	Markov networks
Semi-supervised learning	LDA-based	Backpropagation algorithm
Statistical inference	Gradient ascent	Neural nets
Parameter selection	Q-learning	Risk minimization
EM algorithm	Kernel PCA	Score function
Kernel regression	Bayesian network classifiers	

YELP

Cluster 10 (Deserts/Sweets):

Cake	Pecan pie	Peanut butter cup
Espresso	Indulgence	Custard
Cupcake	Pound cake	Flavored ice cream
Whipped cream	Shake	Chocolate caramel
Brownie	Cinnamon rolls	Ghirardelli
Pumpkin	Oreo	Coconut sorbet
Sundae	Buttercream frosting	

Cluster 24 (Point of Interest Names):

Walgreens	Total Wine	Sushi Roku
FroYo	Pizza Hut	Mc Donalds
Subway	Chinese restaurants	MoJo
Safeway	Wal-Mart	Souper Salad
Denny's	Starbucks	Farm Grill
Costco	Sprouts	TJ Maxx
Target	Wildflower Bread	

Cluster 28 (Traffic/Vehicle related):

Shuttle service	Rental cars	Neon sign
Monorail	Ducati	High traffic
Circle K	Parking lot	Light speed
Ride bikes	Stree parking	Break room
Park area	Garage	Rail line
Terminal to terminal	24/7	Passers by
Honda Civic	Boarding passes	

Cluster 29 (Drinks):

Margarita	Ale	Vodka soda
Iced tea	Shot glass	Cuervo
Sangria	Pinot Grigio	Pint glass
Guinness	Red Bull	Oolong
Martini	Stella	24 oz
Gin	Jamaica	Ladies Night
Draft beer	Drip coffee	

Cluster 39 (Service/Staff/QoS):

Super nice	Barista	Fitting room
Great customer service	Speak English	Great instructors
Hostess	Customer service	Service provider
Cashier	Freaking awesome	Cute waiter
Extremely friendly	Big smiles	Wine guy
Table service	Extremely knowledgable	Personalized service
Kid friendly	Service sucks	

Cluster 33 (Payment and Prices):

Free	$5.50	$5.25
Discount	$9.99	Tip
$1.50	Price tag	lb
Totally worth	Ten dollars	18% gratuity
Unlimited	Resort fee	Automatically added
Bonus	Admission	Half price
Triple	$50.00	

Comparison

As mentioned before, the clustering was done using 10, 20, 40 and 80 centers. It was particularly interesting to see how the granularity of some of the groups increased as the number of centers increased.

For example, for the DBLP data set, when using 10 centers one of them had phrases that could be categorized as "CS Theory" which include things like algorithms, optimization, discrete math; after using 20 centers it was possible to find smaller clusters that were more "specialized", there were clusters that could be identified as "CS Theory | Graphs", "CS Theory | Algorithms", "CS Theory | Optimization".

Some specific examples:
CS Theory | Optimization: non-stationary, cost functions, objective functions, probability distributions, covariance matrix, least squares, design problem, Markov chain, confidence interval, stochastic processes.

CS Theory | Graphs: point sets, planar graphs, shortest path, random graph, directed graphs, spanning trees, matching problem, Nash equilibrium, Voronoi diagram.
CS Theory | Algorithms: SAT solver, well behaved, formally proved, Von Neuman, PSPACE-complete, theorem provers, canonical form, mathematical models, PL/I, Allen, formally verified.

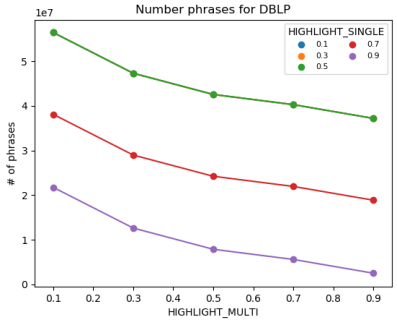
A similar situation was found for the YELP data set. When using 10 centers there was only one cluster that included phrases that identified "Points of Interest", when using 40 centers at least 3 clusters had POIs in it. Unfortunately I was not able to find a pattern to discern what the difference between those 3 clusters were.

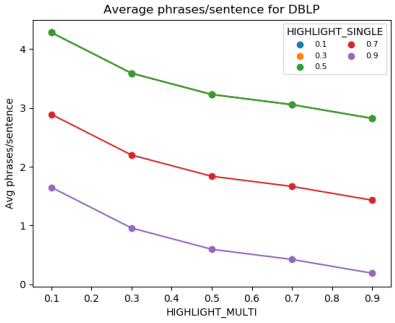
The full data, including some named clusters for each number of centers can be found the GitHub repo.

Parameter analysis

Finally lets look at how the number of segmented phrases and the average number of phrases per sentence changes as the parameters of the segmentation script change. As mentioned earlier in this report 5 values were tried for each of the highlight thresholds resulting in 25 settings. The following plots represent the changes on both metrics as one of the variables changes while keeping the second fixed.

DBLP





YELP