CS 512 Data Mining Principles – Assignment 1

Daniel You (tjyou2@illinois.edu)

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

Dataset	DBLP.300K			
	Single-word		Multi-words	
top-k	collins	kripke	simulated annealing	edit distance
	hebrew	macintosh	quality assurance	anomaly detection
	basque	tiger	neural network	gene expression
	paul	nokia	natural language	belief propagation
	sony	cray	augmented reality	feature selection
	persian	coral	wavelet transform	matrix multiplication
	berlin	ramsey	mixed reality	markov chain
	siemens	swiss	wireless lan	finite state machines
	osgi	gödel	markov chains	virtual reality
	rome	cayley	elliptic curve	neural networks
	gibbs	sdram	turing machines	hash functions
	kim	eiffel	relevance feedback	maximum entropy
	september	michael	vector quantization	reinforcement learning
	michigan	india	timed automata	concurrency control
	thomas	riemannian	decision tree	genetic algorithm
medium-k	webserver	centers	up to 30	an aggregation
	normalised	environments	be reviewed	approach to the problem
	calibration	sign	segmentation via	method to identify
	jb	leadership	better than	develop an efficient
	clfsr	episode	more general	existing ones
	seats	reorganization	n point	techniques for efficient
	conceptually	equivalences	still lack	demonstrate the performance
	circles	mismatch	based on adaptive	can be accessed
	friendliness	vector	researchers who	reason is that
	poly	compatibility	to verify	turn out to be
	equivalents	juntas	some interesting	major drawback of
	extractors	trustees	a preprocessor	approach to reduce
	basketball	decommitment	propose a heuristic	algorithm enables
	multitask	conical	paper describes how	has the advantage
	tailor	facile	lossy compression of	reaction time
bottom-k	22	34	system which has	new approach for the
	could	beforehand	with each other in	be one of the
	taken	000	this paper we describe the	with a series
	used	160	are interested in the	this paper we show
	91	placed	to the identification of	of our approach on
	viz	seems	be used in the	in several different
	05	84	can be used by	be adapted to the
	42	02	find all the	is based on the use
	come	150	to the characteristics of	has been done to
	whole	seem	have designed a	an important role in the
	67	cannot	is that it can	been applied to the
	except	81	as described in	is a need for
	72	concerning	as an example to	that this approach is
	800	indicates	show that they	we discuss some of
	120	indicate	is well known that	been used in the

Dataset	YELP.100K			
	Single-word		Multi-words	
top-k	alas	bourbon	sea salt	soda fountain
	tiki	monterey	papa johns	daily dose
	cuban	cherry	barrio cafe	fettuccine alfredo
	coach	mcdonalds	la fitness	urgent care
	verde	manhattan	cotton candy	del rey
	latin	john	humble pie	vanilla bean
	village	railroad	smoked salmon	cheesecake factory
	capitol	sprinkles	crab puffs	fountain hills
	antipasto	hurricane	chopped liver	collard greens
	queso	zinburger	peanut butter	bikram yoga
	wildflower	national	bell pepper	golden corral
	pierre	macy's	yellow tail	lettuce wraps
	dental	brie	pumpkin porter	dairy queen
	gilbert	hula	chile relleno	los betos
	northern	safeway	pork chop	wild boar
medium-k	extensive	intrigued	to feed	who was visiting
	sweet.the	chair	was incredibly	let him
	envious	providers	we were served	i did enjoy
	babysitting	promote	an important	bothered to
	somebody's	existence	and the selection	i haven't
	elves	round	want a place	was so cool
	good.for	choc	sat near	i really feel
	restless	relive	everything is cooked	and the noodles
	highs	joints	every inch of	was also nice
	clean.i	share	i've spent	had the grilled
	sparkle	uninspired	easy on the	were also pretty
	curves	memorabilia	section of the	rice &
	regrettably	calendars	runs from	to the chandler
	swore	onions	keeping them	removed from the
	fryer	actual	you actually get	within 5
bottom-k	105	concerning	you live in the	had a pretty good
	62	69	and the place has	in the center of
	87	2002	to speak to a	in a place like
	1200	1999	and the salsa was	i really enjoy this
	1960	indicates	be a bit more	i was glad to
	|	59	i guess if you	my husband is a
	97	72	is located in the	for lunch on a
	180	arent	on the list of	out to be a
	\	04	i was glad i	have only been to
	07	causes	is a great place for a	have been coming to
	83	06	is very close to	on one side of
	11.00	01	and the cheese was	is a place you
	98	1950	a visit to the	have a variety of
	05	shed	this location has a	were seated in the
	seeming	350	was the best part of	had to wait about

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

	DBLP.300K	YELP.100K
Number of qualified phrases	956,355	1,560,970
Average number of phrases per sentence	0.677733	0.866286

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

I use nltk.cluster.kmeans.KMeansClusterer and set the number of centers = 20.

DBLP.300K	
Cluster 0 (description: Computer Science) _object-oriented_design_ object-_relational_database_systems_ _backward-compatible_ metaprogramming _database_engine_ _data-flow_ _testing_methodology_ _finite_state_machine_ _net_framework_ _database_schema_ _design_tool_ _abstract_state_machines_ _abstraction_levels_ _source-code_ bigraphical _analysis_tools_ _fault_injection_tool_ _verification_environment_ _machine_language_ _functional_programming_languages_	Cluster 1 (description: Natural Language Processing) _noun-phrase_ _word_sense_ _opinion_mining_ query-focused _cross_lingual_ _clustering_approaches_ _translation_examples_ _semantic_context_ _delimiter_ _text_corpora_ _language_translation_ _hmm_based_ _linguistic_analysis_ vocabularies _parallel_corpus_ _computer-assisted_translation_ _textual_data_ enriching _automatically_extracts_ transliterations
Cluster 2 (description: Machine Learning) _optimization_algorithms_ _multiagent_reinforcement_learning_ _hybrid_methods_ _sequence_alignments_ annealing em-like _based_adaptive_ _semi-markov_ _continuous_state_ _approximate_inference_ _multiagent_learning_ _multiclass_ _nearest_neighbor_queries_ _hierarchical_bayesian_ _propagation_method_ filter _curse_of_dimensionality_ _clustering_techniques_ _feature_weights_ mlps	Cluster 3 (description: Computer Architecture) _fpga_devices_ _processor_arrays_ dma 8-node _heterogeneous_multiprocessors_ _back_end_ microprogrammed synthesizer _soc_designs_ _based_systems_ _workload_characterization_ _bus-based_ micro-processor _dynamically-scheduled_ _parallel_machine_ _super-scalar_ _highly-parallel_ non-pipelined _single_processor_ _tasks_scheduling_

YELP.100K	
Cluster 0 (description: Time)	Cluster 1 (description: Feelings)
_friday_afternoon_	surprises
_pretty_late_	_forgettable_
morning's	_low_rating_
_late_night_dinner_	unpredictable
_sunday_nights_	_letdown_
_black_friday_	horrible
_tuesday_morning_	_dislike_
_friends_birthday_	critical
winter	_lame_
_soft_opening_	_decent_breakfast_
mid-afternoon	underrated
after-hours	uninspired
_mothers_day_	wasteful
_business_lunch_	_five-star_
_busy_nights_	_memorable_experience_
_weekday_night_	_highly_doubt_
weekends	_food_poisoning_
_open_til_	_high_prices_
_weekend_night_	_horrible_food_
workday	_biggest_issue_
Cluster 5 (description: Desserts and Drinks)	Cluster 10 (description: Opinions)
_sour_cherry_	leftovers
pastry	_casual_dining_
_chocolate_torte_	_average_prices_
_red_bean_paste_	_fairly_decent_
_homemade_ice_cream_	_decent_quality_
smooth	_thumbs_up_
milkshake	_yummy_food_
caramels	_awesome_
_kettle_corn_	_decently_priced_
espresso	_friendly_atmosphere_
tea	_dirt_cheap_
_chocolate_milk_	_attentive_service_
shortcake	_reasonable_prices_
pudding	_cheap_meal_
_strawberry_sauce_	_outstanding_food_
_raspberry_sorbet_	_neighborhood_place_
_lemon_tart_	_casual_atmosphere_
gingerbread	_decent_service_
godiva	_tastes_amazing_
_waffle_cone_	well-versed

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.

(1) Draw the number of phrases versus HIGHLIGH_THRESHOLD curve for both multi-words and single word phrases returned by AutoPhrase.

Fix HIGHLIGHT_MULTI to 1.0 and adjust HIGHLIGHT_SINGLE from 0.0 to 1.0 in 0.2 increments.

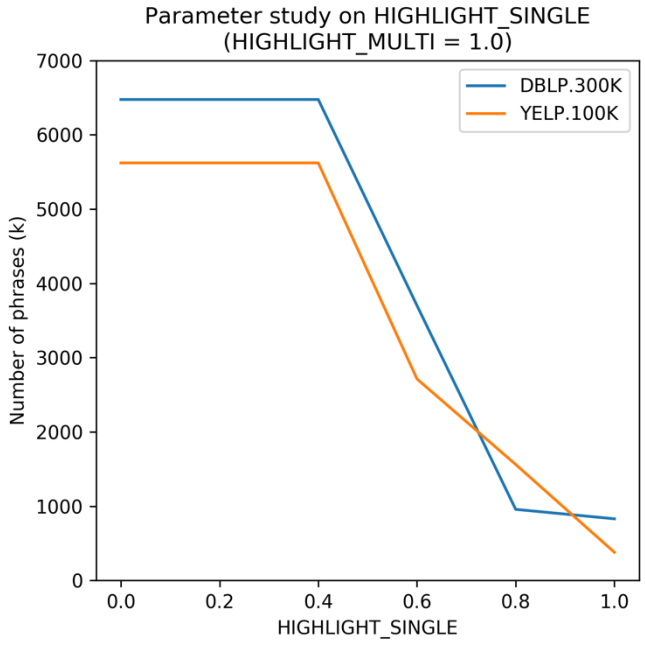
DBLP.300K						
HIGHLIGHT_SINGLE	0.0	0.2	0.4	0.6	0.8	1.0
HIGHLIGHT_MULTI	1.0	1.0	1.0	1.0	1.0	1.0
Number of Phrases	6,474,506	6,474,506	6,474,506	3,700,344	956,355	829,164

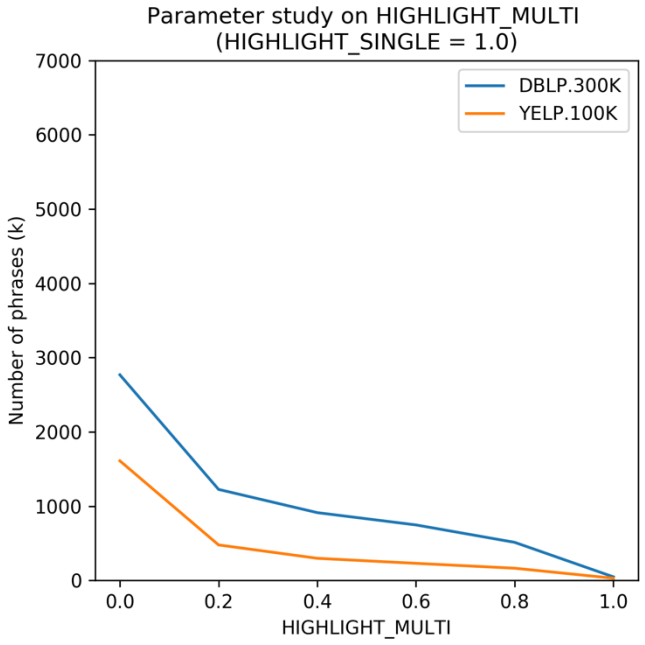
YELP.100K						
HIGHLIGHT_SINGLE	0.0	0.2	0.4	0.6	0.8	1.0
HIGHLIGHT_MULTI	1.0	1.0	1.0	1.0	1.0	1.0
Number of Phrases	5,620,537	5,620,537	5,620,537	2,712,355	1,560,970	378,531

Fix HIGHLIGHT_SINGLE to 1.0 and adjust HIGHLIGHT_MULTI from 0.0 to 1.0 in 0.2 increments.

DBLP.300K						
HIGHLIGHT_SINGLE	1.0	1.0	1.0	1.0	1.0	1.0
HIGHLIGHT_MULTI	0.0	0.2	0.4	0.6	0.8	1.0
Number of Phrases	2,767,982	1,224,934	912,284	746,795	512,180	47,565

YELP.100K						
HIGHLIGHT_SINGLE	1.0	1.0	1.0	1.0	1.0	1.0
HIGHLIGHT_MULTI	0.0	0.2	0.4	0.6	0.8	1.0
Number of Phrases	1,609,528	477,466	297,388	229,658	163,845	28,895





(2) By setting number of centers, you can get phrase clusters in different granularity. Show some representative clusters and 10 words in each cluster for different granularity. (e.g. k = 5, 10, 25)

DBLP.300K		
k= 5	k = 10	k = 25
Cluster 0 (description:) theweb domain-specific _relational_data_base_ _knowledge_structures_ _egee_ _distributed_hash_tables_ _language-based_ _separation_of_duty_ neues konturen	Cluster 1 (description:) _web_communities_ _inverted_file_ _precision/recall_ pedigree _holistic_twig_ _text_classifier_ _learning_phase_ _keyphrase_extraction_ _quad_tree_ _bilingual_dictionaries_	Cluster 0 (description: Computer Science) _view_integration_ _escher_ _state_space_analysis_ _kit_ marte _standard_uml_ erts _owl-s_ _domain_specific_language_ sbql
Cluster 1 (description:) _hemoglobin_ garch _single_frame_ bias tagging btfs _document_space_ inflection _hyper_-spectral _newton_method_	Cluster 3 (description:) _reed-solomon_ _instruction-set_processors_ _automatic_test_generation_ sinking 8-way hand-optimized _qwerty_-like _voltage-scaling_ _systematic_design_ performance-complexity	Cluster 3 (description: Computer Architecture) _fpga_devices_ _processor_arrays_ dma 8-node x-tree _heterogeneous_multiprocessors_ _back_end_ _fir_ spe openrisc
Cluster 2 (description:) mdst x_i _aggregate_signatures_ push-out _job_shop_ _forgery_attack_ halts ssat prespecified _joint_source_	Cluster 4 (description:) _obstacle-avoiding_ nlog2 parentheses _outerplanar_graph_ _algebraic_number_field_ lim out-branching _expected_distortion_ hyperheuristic parameter	Cluster 6 (description:) _large_spatial_ _backscatter_ malignancy _gene_products_ pumps _expression_patterns_ _rna_secondary_structure_ per- electromyographic acute
Cluster 4 (description:) auctioneer/buyer _eye_tracking_ humans collected interview communicate _physical_objects_ _private_information_ weeks uncover	Cluster 7 (description:) _sdb_ kooperative ddc eng 5- fréquents dissections superstrings linearen _asp	Cluster 9 (description:) rserpool _dynamic_service_ _mobile-agent_based_ _management_issues_ pacts ome _educational_hypermedia_ inter-organization http-based database-driven

YELP.100K		
k= 5	k = 10	k = 25
Cluster 0 (description:) deterrent _great_pho_ exemplary _neapolitan_ _loose_leaf_ _previously_mentioned_ vastly _aloha_kitchen_ hh _pastys_	Cluster 0 (description: Time) _returned_home_ _tasty_meal_ extending _valentines_day_ _mon_-_thurs_ _friday_ tuesday luci's _friday_morning_ _seahawks_	Cluster 0 (description: Time) _friday_afternoon_ _sunday_nights_ _black_friday_ _patrick's_day_ _tuesday_morning_ _friends_birthday_ _mothers_day_ _busy_nights_ _weekend_night_ _lunch_hours_
Cluster 1 (description:) corroded leaky _classical_music_ _particle_board_ _construction_ blare _picnic_bench_ jugs _over-run_ _tobacco_	Cluster 2 (description:) _great_breakfast_ well- 83 _ipa's_ _foo-foo_ _organic_produce_ depending service- solid mocha's	Cluster 3 (description:) _middle_aged_ cliches woos relatives _nhl_ _grandparents_ spectators watch sex _business_travelers_
Cluster 3 (description: Feelings) _valentines_day_ hears unenthusiastic affiliation _handgun_ 5-minute sharpening _flaunt_ _low-cost_ growlers	Cluster 6 (description: Foods) _beer_glass_ _chocolate_raspberry_ cinnamony _navy_ rectangular avacados tangy _tender_chicken_ _red_wine_ icey	Cluster 2 (description:) hick knowingly optimist naw _woo_ pagers misread accomplished stephy fryers
Cluster 4 (description: Foods) _candy_cane_ salad _jollof_rice_ appys tsatziki rehydrated quirkiness jimmies greenbeans mix-n-_match_	Cluster 9 (description: Menus) trio baked wifes dol _sizzling_rice_soup_ _holy_crap_ _marinated_beef_ _pulled_chicken_ _wontons_ _croissant_sandwich_	Cluster 6 (description:) domed embossed _stainless_steel_ fluorescent _large_bar_ equipped beaches courtyard _mini_bar_ shower