

# NATURAL LANGUAGE PROCESSING ON STUDENT REVIEWS OF CLARK



CLARK  
UNIVERSITY

7/18/2018

Ralitza Dinesh Mondal



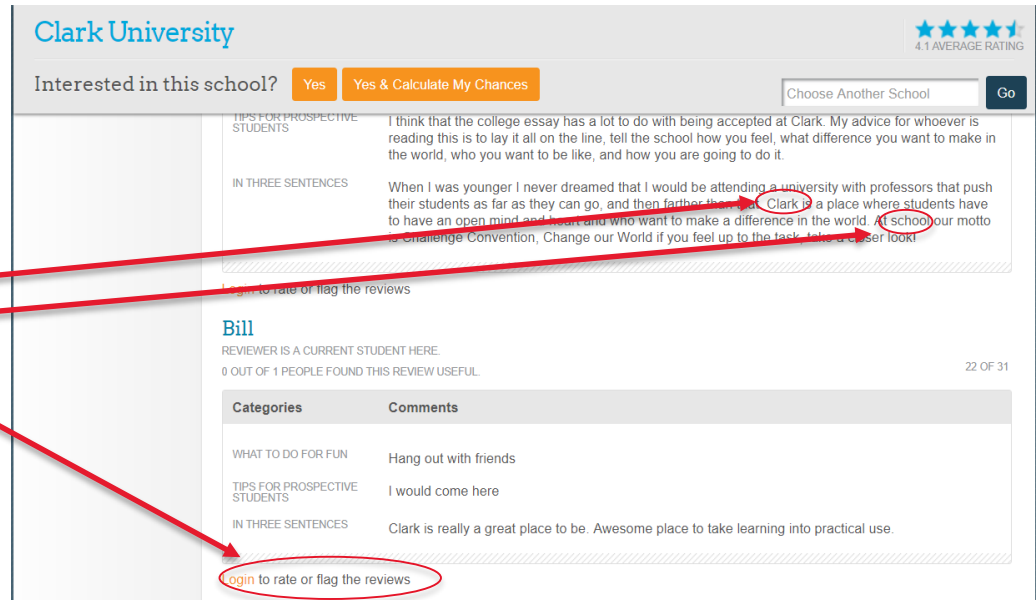
# Introduction

In Total: 1329 Reviews were scraped from four websites:

- <https://www.niche.com/colleges/clark-university/>
- <https://www.cappex.com/colleges/Clark-University/reviews>
- <https://www.unigo.com/colleges/clark-university>
- <https://www.zhihu.com/question/43661439> (Chinese website to get review from an international perspective)

# Use of R programming language

- Cleaning the text
  - Remove numbers
  - None and missing comments
  - Remove punctuation
  - Remove unnecessary noise words and common words

A screenshot of a Clark University review page. The page has a header with the Clark University logo and a 4.1 average rating. Below the header, there are buttons for "Interested in this school?", "Yes", "Yes & Calculate My Chances", and a "Choose Another School" button. The main content area contains a review by a user named "Bill". The review is divided into sections: "TIPS FOR PROSPECTIVE STUDENTS", "IN THREE SENTENCES", and "WHAT TO DO FOR FUN". The review text is: "I think that the college essay has a lot to do with being accepted at Clark. My advice for whoever is reading this is to lay it all on the line, tell the school how you feel, what difference you want to make in the world, who you want to be like, and how you are going to do it." The "IN THREE SENTENCES" section contains: "When I was younger I never dreamed that I would be attending a university with professors that push their students as far as they can go, and then farther than that. Clark is a place where students have to have an open mind and look and who want to make a difference in the world. At school our motto is Challenge Convention, Change our World if you feel up to the task. Take a closer look!" The "WHAT TO DO FOR FUN" section contains: "Hang out with friends". The review is rated 0 out of 1 people found this review useful. At the bottom of the review, there is a link to "login to rate or flag the reviews". Red arrows point from the text "Remove unnecessary noise words and common words" to the review text, and a red circle highlights the "login to rate or flag the reviews" link.

# Words are tokenized and frequency of words



docs	text1	text2	text3	text4	text5	text6	text7	text8	text9	text10	text11	text12	text13	text14	text15
eatres	0.07692308	0.00000000	0	0	0	0	0	0.11111111	0	0	0.00000000	0	0.00000000	0.00000000	0
spree	0.15384615	0.00000000	0	0	0	0	0	0.22222222	0	0	0.00000000	0	0.00000000	0.00000000	0
day	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
class	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
cancel	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
everyon	0.07692308	0.00000000	0	0	0	0	0	0.11111111	0	0	0.00000000	0	0.00000000	0.00000000	0
drink	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
univers	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.11764706	0	0.16666667	0.00000000	0
sponsor	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
pet	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
zoo	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
bounci	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
hous	0.07692308	0.00000000	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.07142857	0
explor	0.00000000	0.1428571	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0
worcest	0.00000000	0.1428571	0	0	0	0	0	0.00000000	0	0	0.05882353	0	0.00000000	0.07142857	0
take	0.00000000	0.1428571	0	0	0	0	0	0.00000000	0	0	0.00000000	0	0.00000000	0.00000000	0

Each words is tokenized, creating a bag of words.

Each review is a document and stored under the name of “text”

This table shows the frequency of specific word in all the text reviews.

# Giving weight to the frequency of words



spree	day	class	cancel	everyon	drink	univers	sponsor	pet	zoo	bounci	hous	explor	worcest	take
2.0443437	1.3911312	0.9276253	2.1692825	1.2370343	1.6464037	0.9166991	2.6464037	3.1235250	3.1235250	2.8224950	1.2370343	1.7811023	0.9805102	1.1149248

From the previous matrix of 15 tokens:

Each tokenized words is given a weight where the higher value represents that rarer words are given a higher importance and chances are higher that they might contain important semantic meaning.

$$w_{ij} = tf_{ij} * \log_2(N/n),$$

So for the weight to be high, N is the number of texts and the n is the number of document that have that token. So lower n gives higher weight.



# Reduced dimensions by SVD

From the previous matrix of 15 tokens:

We further reduced the dimension by using singular value decomposition. This finds a basis for documents using concept vectors and the finding a basis for concept using word tokens as vectors. Decompose of a matrix into two simple matrices where one matrix is the relationship between words and concepts and the other is the relationship between the concept and the documents

```
> lsa_model$sv[1:15] [1] 3.631207  
3.076223 2.429687 2.348920 2.323238  
2.294575 2.198737 2.189959 2.168683  
2.079944 1.881995 1.839868 1.831018  
1.823439 1.794445
```

Why we do this?

Gives me the n-best-rank approximation to the original matrix. I can create a reduced dimension of the original matrix preserving the original meaning.

# Model Summary for first 15

```
> lsa_model$summary
```

	topic	label_1	coherence	prevalence	top_terms
t_1	t_1	none	0.180	3.812	none, soror, frat, wonder, thing, peopl, go, frustrat, haven, fratern
t_2	t_2	spree	0.599	-3.384	none, risk, declin, micro, blog, couldn't, flesh, downturn, repeal, nightfal
t_3	t_3	republican	0.558	3.311	republican, affili, kind, conserv, peopl, origin, nation, divers, disappoint, polit
t_4	t_4	beat	0.221	6.050	beat, anni, brunch, greek, life, great, campus, woooo, good, best
t_5	t_5	greek	0.190	35.800	brunch, anni, greek, life, woooo, great, campus, best, food, place
t_6	t_6	greek	0.065	48.634	greek, life, woooo, great, campus, best, food, peopl, strict, good
t_7	t_7	greek	0.114	-6.995	greek, life, exist, recogn, brunch, anni, great, negat, big, fraternati
t_8	t_8	usermay	0.325	3.471	usermay, 2010food, nich, sure, introvert, extrovert, shi, possit, make, attitud
t_9	t_9	strict	0.029	27.620	strict, campus, good, peopl, lenient, safe, lot, love, student, help
t_10	t_10	spree	0.075	9.939	food, love, peopl, lot, place, great, mani, salvadorian, campus, help
t_11	t_11	food	0.286	-6.843	salvadorian, pupusa, el, food, vietnames, great, love, place, cafeteria, strict
t_12	t_12	sure	0.087	-4.977	sure, safe, danger, salvadorian, make, pupusa, el, campus, vietnames, great
t_13	t_13	psycholog	0.313	-0.205	psycholog, geographi, depart, major, program, prestigi, professor, help, freud, peopl
t_14	t_14	psycholog	0.047	-2.935	psycholog, sure, worth, salvadorian, pupusa, el, depart, geographi, make, prestigi
t_15	t_15	danger	0.133	-1.090	danger, psycholog, surround, neighborhood, safe, pretti, salvadorian, pupusa, el, worth

Coherence: It is the correlation between the top terms in the specific topic in the semantic space

Prevalence: It is the covariance between the review frequency and the topic.

Negative value meaning that as the document size grew, the topic was with that combination of top-terms were seen less and less. The positive value shows opposite meaning.



# Top Topics: 1. Fraternity

t_1	Fraternity	0.18	3.812	none, soror, frat, wonder, thing, peopl, go, frustrat, haven, fratern
t_56	Fraternity	0.158	1.573	soror, frat, expans, stingi, money, frustrat, haven, fratern, progress, price
t_7	Fraternity	0.114	-6.995	greek, life, exist, recogn, brunch, anni, great, negat, big, fraternati
t_6	greek	0.065	48.634	greek, life, woooo, great, campus, best, food, peopl, strict, good
t_5	greek	0.19	35.8	brunch, anni, greek, life, woooo, great, campus, best, food, place

This is the first most common topic talked about throughout the review with a prevalence of 48.634. The reviews highly mentions the absence of fraternity and t\_56 stands out with a negative word like “frustrate”.



## Top Topics: 2. Community

t_9	Communit	0.029	27.62	strict, campus, good, peopl, lenient, safe, lot, love, student, help		
t_10	Communit	0.075	9.939	food, love, peopl, lot, place, great, mani, salvadorian, campus, help		
t_37	Communit	0.011	1.483	come, health, place, awesom, freud, challeng, divers, perfect, mind, open		
t_34	Communit	0.015	0.956	place, kind, vietnames, perfect, club, lot, worcest, exact, park, everyon		
t_81	Communit	0.061	0.637	neighborhood, surround, bad, im, parti, gym, unaccept, wifi, accept, larg		
t_80	Communit	0.032	0.438	exact, right, swim, experi, neighborhood, program, surround, centr, communiti, hang		
t_28	Communit	0.127	-1.046	kind, commit, good, liber, scholarship, gave, valu, awesom, social, hall		
t_64	Communit	0.034	-1.107	toward, challeng, progress, visit, similar, look, friend, hang, club, lgbt		
t_65	Communit	0.016	-1.38	vietnames, great, visit, similar, parti, im, communiti, school, other, let		
t_23	Communit	0.023	-3.01	help, involv, servic, smoke, weed, career, peopl, love, centr, xenophob		
t_74	Communit	0.033	-3.087	much, drug, help, interest, lot, know, weed, vietnames, opportun, toward		
t_18	Communit	-0.009	-3.158	love, food, im, clean, communiti, now, toward, lgbt, clarki, beauti		

Positive aspects talked about the campus are being safe and **people are lenient and student love and help**. People talk about great Salvadorian place around campus that has **great food**. They also talk about challenge convention and the **diversity** and open mind at Clark and the **Vietnamese community** around Clark. The lesser talked positive aspects were Clark having LGBT community and parties around campus. There are less prevalent **negative aspect** mentioning the **poor wifi** and **unsafe neighbourhood**. They talk about xenophobia and smoking of weed.



## Top Topics: 3.Conservatives

t_76	conserv	0.251	1.876	peppercorn, polit, array, espec, conserv, club, much, help, im, concious		
t_99	conserv	-0.008	-0.039	fun, conserv, excit, drug, help, athlet, part, communiti, vietnames, frustrat		
t_75	conserv	0.1	-0.516	polit, concious, bodi, blackston, shop, activ, exact, bus, let, etc		
t_3	Conservati	0.558	3.311	republican, affili, kind, conserv, peopl, origin, nation, divers, disappoint, polit		
t_82	Conservati	-0.01	1.116	right, exact, polic, conserv, hang, stingi, polit, cool, wish, option		
t_83	Conservati	0.002	-0.126	bad, im, everyon, lot, psych, conserv, much, unaccept, now, option		
t_77	Conservati	0.046	-0.158	polit, much, class, concious, perfect, conserv, bodi, weirdo, everyon, gym		

Conservative: This is the third most talked topic about throughout the reviews. Mostly negative aspects, republicans are disappointed and the political view is not acceptable. Though Clark being a diverse school, the people are “stingi” about conservatives. Few reviews also say that Clark is not politically conscious and have perfect conservative student body.

# Topics that stand out with unique words



Food: This topic is less prevalent but they talk about important keywords for food such as el Salvadorian, Pupusa and Vietnamese food emphasizing these types of food are probably the most popular around. The cafeteria is also mentioned being strict, probably on dietary needs.

Psychology: This is one of the less prevalent topics but when reviews do mention, they talk about psychology and geography being important majors and prestigious professors teaching those.

Smoke: This is also one of the less prevalent topics but the appearance of this topic does stand out as people mention there is a weed and drinking and increase in the same list of top terms.

Food had a more neutral aspect discussing the most popular food around campus with unique words like “el salvador”. Psychology stands out for the unique mention of the word “major”, “prestige” and “professor”. And, smoke, stands with because of the words “always” and “dorm” related to it.



## Other moderately important themes

**Activities:** The highest prevalent words were: bus, consist, free, weekend, care, help, great, community. People talked the most about the free weekend buses to the mall and the buses being consistent. Along with student clubs progress. The lesser talked about aspects about activities were having activities in the park and fun people around campus.

**Aid:** The most coherent words that stood out in the entire document were: help, aid, financial, service, career, scholarship, health, qualify, centre, good. So, people mostly talked about the great financial aid and service and that helps find good career help probably through the leap center.

**Aesthetic:** through the coherency and prevalence of the words, people find the facilities clean and beautiful and accessible. They also talked about the accessibility of the dorm and houses around campus.



## Other moderately important themes

**Campus:** The most prevalent words that came about with high coherence were: party, room, spacious, club, visit. So people talked about the having parties in the dorm rooms because of its spacious size. They also talk about having easy atmosphere to hang out with friends in the spacious dorm rooms.

**Campus Life:** The most prevalent words seen with good coherence were: “go, gym, restaurant, focus”. People talk about the going to the gym and that helps them focus. They also talk about classmate and working and having fun can be frustrating to manage but having progress can be hard. Overall, having job and focusing can be hard and frustrating but the gym helps them focus. Another aspect which is not that prevalent but is mentioned, is the convenience of Clark escort.

**Career:** The most prevalent words with good coherence seen were: “centre, career, accept, leep, prepare, ensure, well”. So people are mostly talking about how the leep center is helping them to prepare well for their career.



## Other moderately important themes

Dorm: They talk about athlete needing more dorm and wifi being unacceptable and that there is the presence of drug always. The lesser talked aspects were positive, hanging around the dorm and resources are always available and cheap.

Environment: People mostly talk about the cold weather throughout the year and how it makes them sad. The presence of LGBT and Vietnamese communities is also mentioned.

Sport: This topic is moderately talked about throughout the reviews, and mostly positive sentiments. There is a good spirit for sport and the teams are safe and there is access to gym. There are also options to play in the cold. They also mention financial aid for athletes and people also mention of sports like football and swimming.

Academics: The top terms that stood out were great, open, mind, focus and professor. So people are generally talking about being open minded and focus and the professor's quality of teaching.

# Insight on the reviews talked about career and leep center from most recent



College FreshmanMay 12 2016Overall Experience  
Very helpful LEEP centre and Career Services

College FreshmanJan 13 2016Health & Safety

The leep center seems very helpful but it can be overwhelming to navigate and understand how the system works. However, once you get the hang of it, it can be essential to paving your future

College JuniorFeb 28 2016Overall Experience

Their is no alumni networking. And their is a huge lack of diversity in terms of employees being hired. Career services tries their best to help, but it also lacks the ability to cater to every student.

College StudentMay 13 2016Overall Experience

Career and internship opportunities are plenty for liberal arts majors but very little for science majors. Very enclosed within Worcester as well.

College SophomoreDec 30 2015Overall Experience  
Career Services is not helpful at all.

College FreshmanNov 18 2013Student Life

The Leep Center is a good resource that the school provides with assisting students with academics, registration, study abroad and much more.

College SophomoreOct 10 2013Campus

Most students still don't understand what LEEP is but it a unique and extremely helpful program available to all Clark students. At the LEEP center they help students who want to apply what they're learning to the real world through internships, study opportunities and in helping to find careers for when students graduate. At Clark you know those working there want you to succeed and they will give you whatever resources you need to do whatever you want to do.

Visit the Leap

for services such as: career services, writing assistance, academic advising, and other valuable services.