Paper SESUG 246-2018

Text Mining to Predict College Admission Trends

Shashikant Chebrolu, Oklahoma State University

Abstract

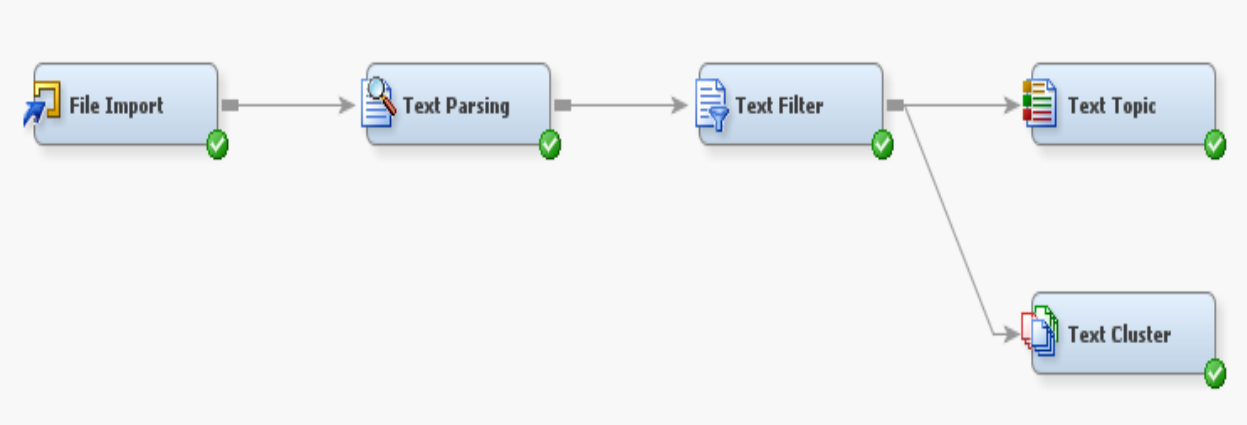
Most higher education institution receives numerous applications for college admission. Out of those, some lucky ones get an offer of admission to the college. But from the pool of students who are given the admissions, there are many those who do not accept the offer. Due to this, a student who's deserving and willing to join a particular college may lose out an opportunity to do so.

This paper aims at finding the link between the students who accept an admission offer and their prior interactions with college over emails to determine if there is a pattern that can be used to predict acceptance and suggest ways to increase the conversion rate. The data used has been provided by OSU's Institutional Research & Information Management. The dataset contains all the communications between the student being offered the admission and the university for the past two years. SAS Enterprise Miner 14.2 has been used for conducting text analysis in this project.

The goal of the project is to incorporate the text analytics results to identify the topics that are identified in student initiated emails with the university so more accurate decision making can be performed during the admission process.

Methodology

The datasets used have been acquired from OSU's Institutional Research & Information Management. The primary data set contains all the interactions to and from a student who applied for a position. This data was matched with other data sets containing the enrollment information of students. This facilitated in determining whether the student being offered a seat accepted the offer or not. Other data cleaning and preparation techniques like joins were done using JMP. Once the final dataset was ready, text parsing was done using online updated dictionary. After parsing the text, data filtration was done followed by clustering and text topic building.



Data Preparation

Primary dataset contained all the emails exchanged between prospective students and the university.

Semester wise enrollment data containing the demographics of the students was available in a separate dataset. This dataset contained the details of all the students who accepted an admission offer.

The primary dataset was merged with semester wise data to determine the students who accepted an offer. A target column was created to represent this set of students indicating 1's for those who accept an offer and 0's for those who don't.

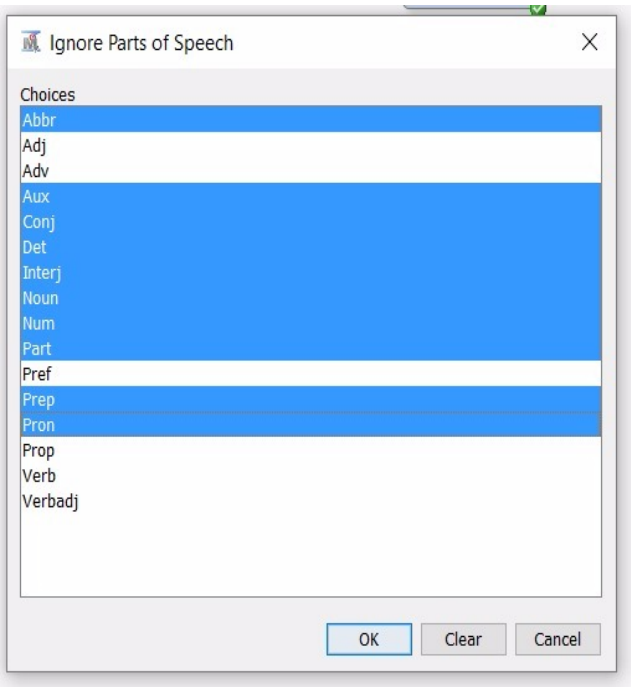
Data cleaning involved removing all the email addresses from the body of the text to reduce misrepresentation.

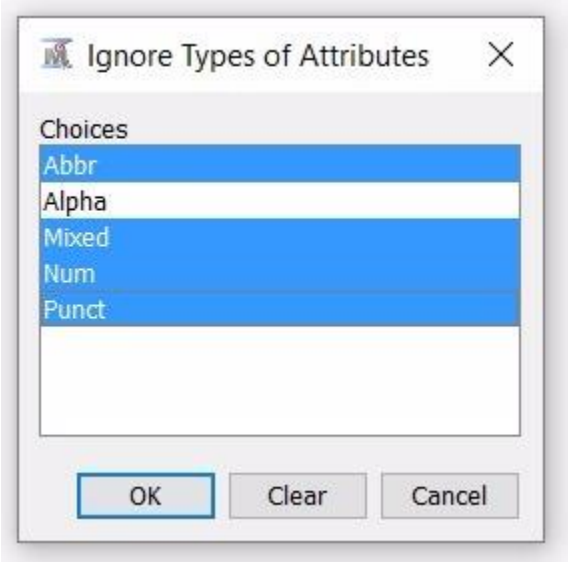
Only the conversations initiated by the student were considered for the analysis.

Data Processing for Text Analytics

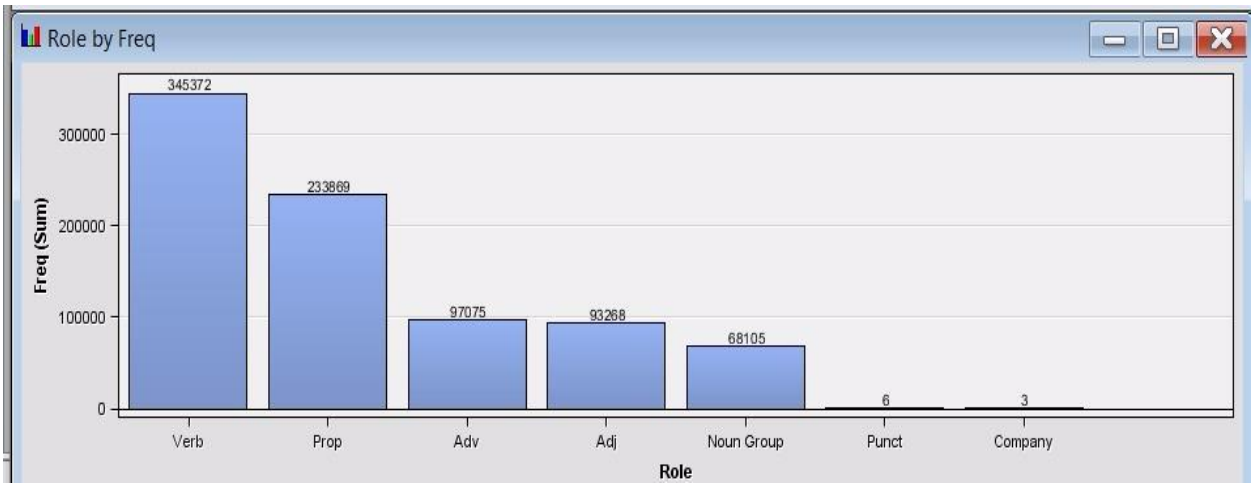
Text Parsing: Text parsing node of SAS Enterprise Miner was used to clean the data. Certain parts of speech were ignored like Abbreviations, Auxiliary, Conjunction, Determinant, Interjection, Noun, Number, Preposition and Pronoun to remove repetitive words like OSU, Stillwater and so on.

Certain Types of Attributes were ignored namely Abbreviation, Mixed, Number & Punctuation





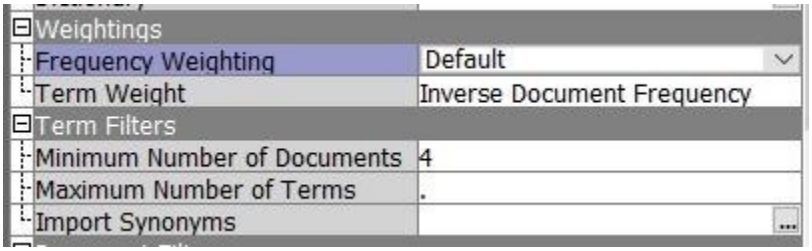
Stop words like all, and, any etc were removed.



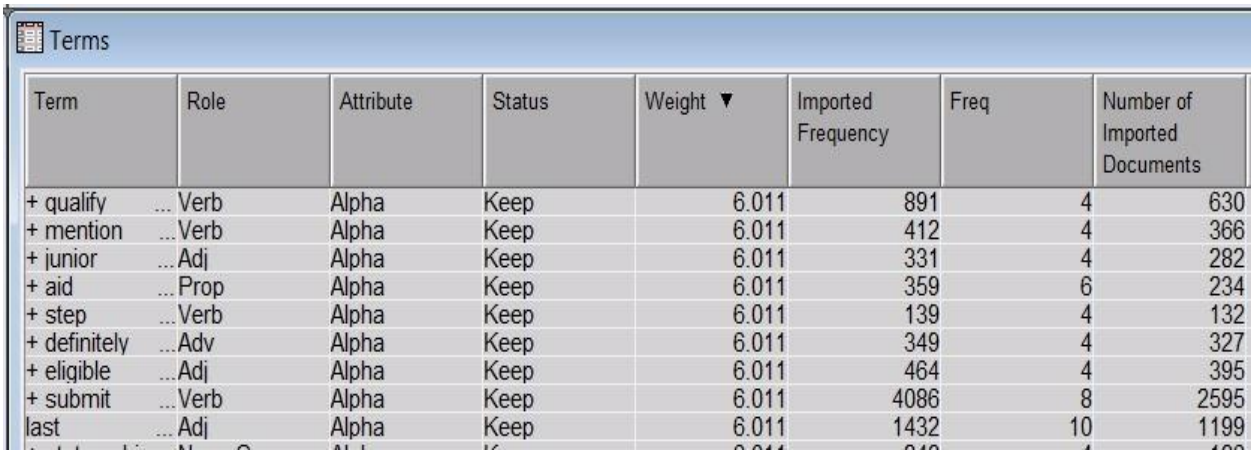
After running the text parsing node, these were the proportion of the various parts of speech that could be used for further analysis.

Text Filtering: Text filtering node was used to assign weightage to the words. Words with high weightage help in identifying the documents. Words with low weightage were dropped and those with considerable weightage were kept for further analysis.

Weightages were calculated based on the settings mentioned below



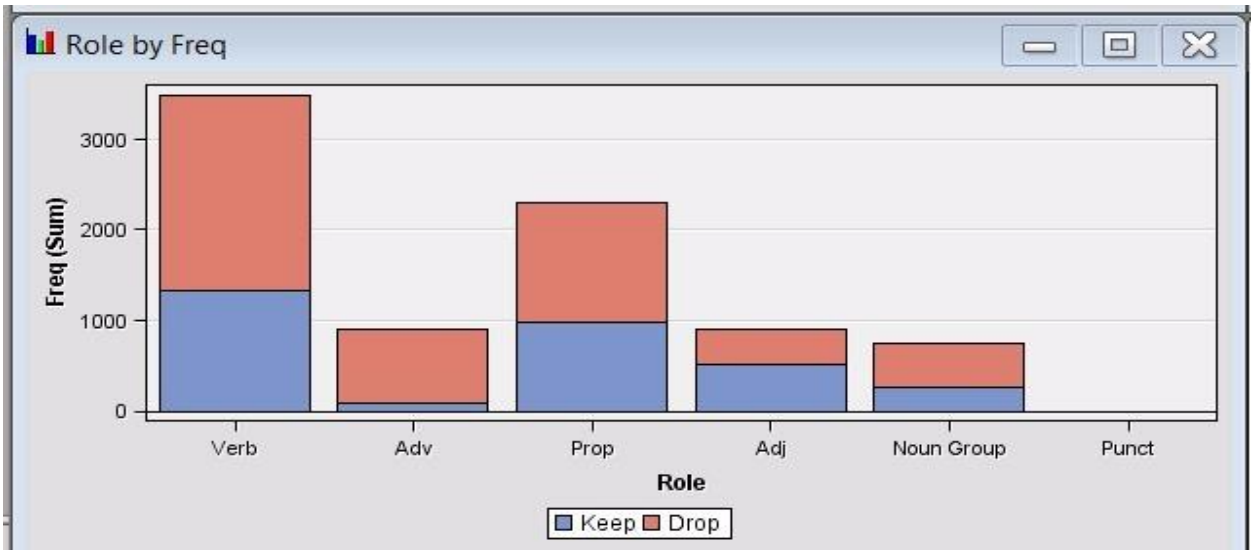
The following are the words with a high weightage.



Term	Role	Attribute	Status	Weight ▼	Imported Frequency	Freq	Number of Imported Documents
+ qualify	Verb	Alpha	Keep	6.011	891	4	630
+ mention	Verb	Alpha	Keep	6.011	412	4	366
+ junior	Adj	Alpha	Keep	6.011	331	4	282
+ aid	Prop	Alpha	Keep	6.011	359	6	234
+ step	Verb	Alpha	Keep	6.011	139	4	132
+ definitely	Adv	Alpha	Keep	6.011	349	4	327
+ eligible	Adj	Alpha	Keep	6.011	464	4	395
+ submit	Verb	Alpha	Keep	6.011	4086	8	2595
last	Adj	Alpha	Keep	6.011	1432	10	1199

Words that occur in less than four documents were also dropped.

Based on the weightages and frequencies of the terms, words were filtered out as per the table below. The terms that were retained were used for further analysis.



Text Topic: Text topics were identified to gain insights on what topics the students are most concerned about. This information can be used by the university to it's advantage by addressing those concerns even before the student initiates.

The text topic node was run using the default settings. The results seemed to be biased since a large number of nouns (for example admission counsellors of OSU) appeared in the results even after being filtered out in the Text Filter node.

Due to these biases, I had to go back to Text Filter node to manually filter out the words which were influencing the results negatively using the Interactive Filter Viewer option within the Text Filter node. This process had to be repeated a few times to get the text topics that made sense.

Term Topics	
Number of Single-term Topics	0
Learned Topics	
Number of Multi-term Topics	7
Correlated Topics	No
Results	
Topic Viewer	

The above table shows the settings used for running the text topic node.

Results

Based on the above analysis, the following text topics were identified to be together:

Topic ID	Topic
1	+submit,+receive,leadership resume,+high,+essay question
2	financial,financial,+aid,+financial aid,hanis
3	+admissions,+counselor,+east,student,union
4	+meet,+osu,+attend,+interest,+love
5	transfer,+flower,mound,+statistics,north
6	weston,sat,gpa,+academic,+high
7	+defer,whaley,+madi,+cancel,+register

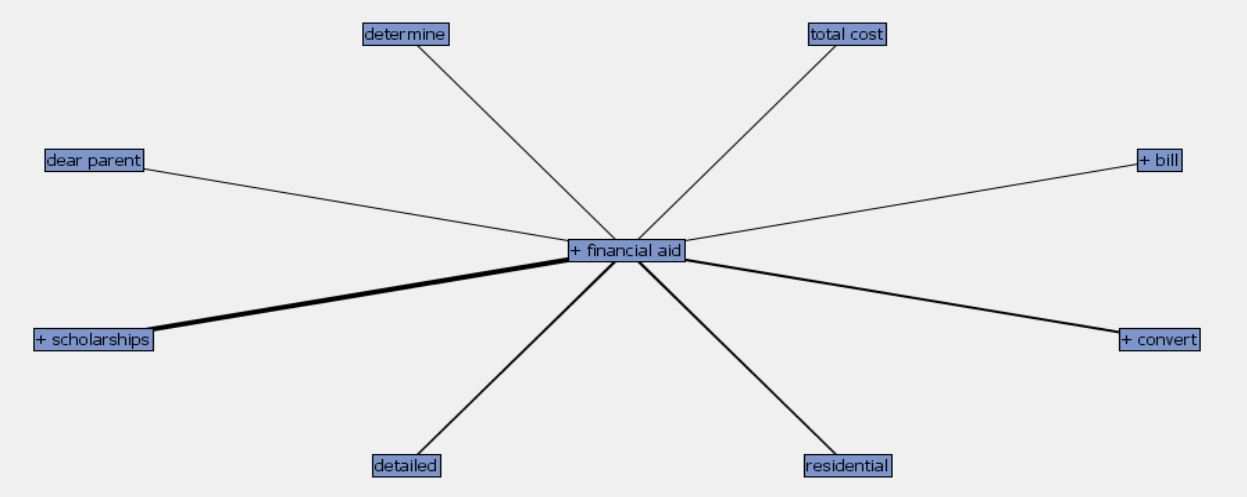
These are the clusters of words that occur most frequently together.

Topic 1: submit, receive, leadership resume, high, essay question

Here, the student might confirm the university that they have submitted the required documents and how to proceed from there. Nothing conclusive other than that can be found from this cluster.

Topic 2: financial aid

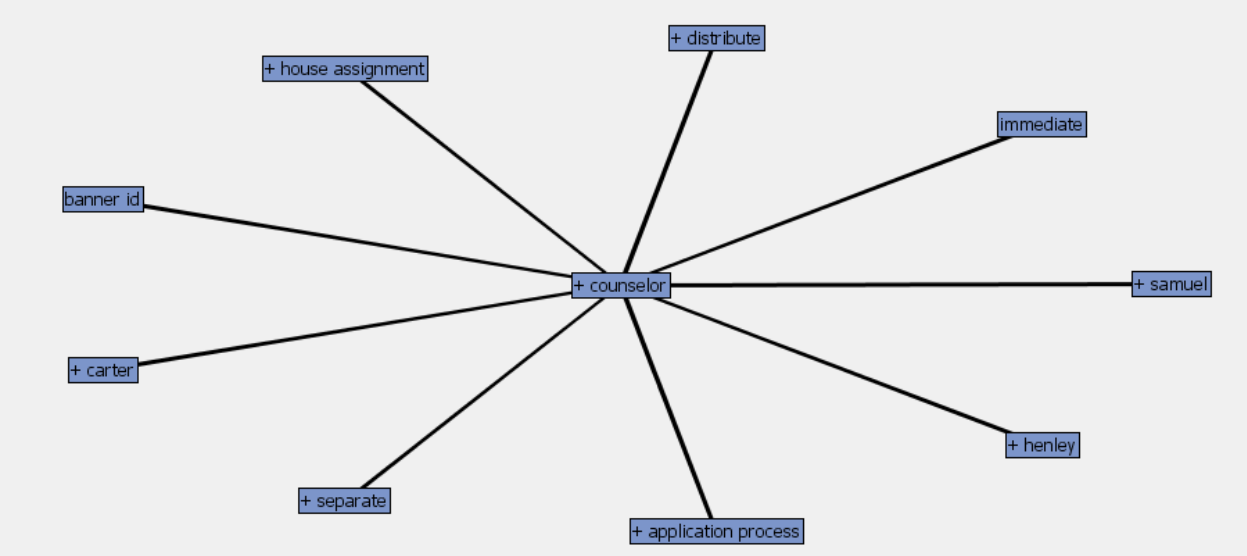
The student is curious about receiving some kind of financial aid.



From the concept link diagram, these are the words that are strongly associated with the term financial aid. The university can address these concerns beforehand for a smoother admission process to the student. It can email important links and contact information regarding financial aid so that it would be helpful for the student.

Topic 3: admissions, counselor, east, student, union

The student is querying about admissions counselor most probably to know about their respective academic counsellor so that they can understand the admission process better.



The concept link diagram for counselor shows the words strongly associated with it. These words show what kind of questions the students ask their respective academic counselor.

Topic 4: meet, OSU, attend, interest, love

This shows that the student is interested to attend OSU. Nothing conclusive other than the student's interest can be observed from this cluster.

Topic 5: transfer, flower, mound, statistics, north

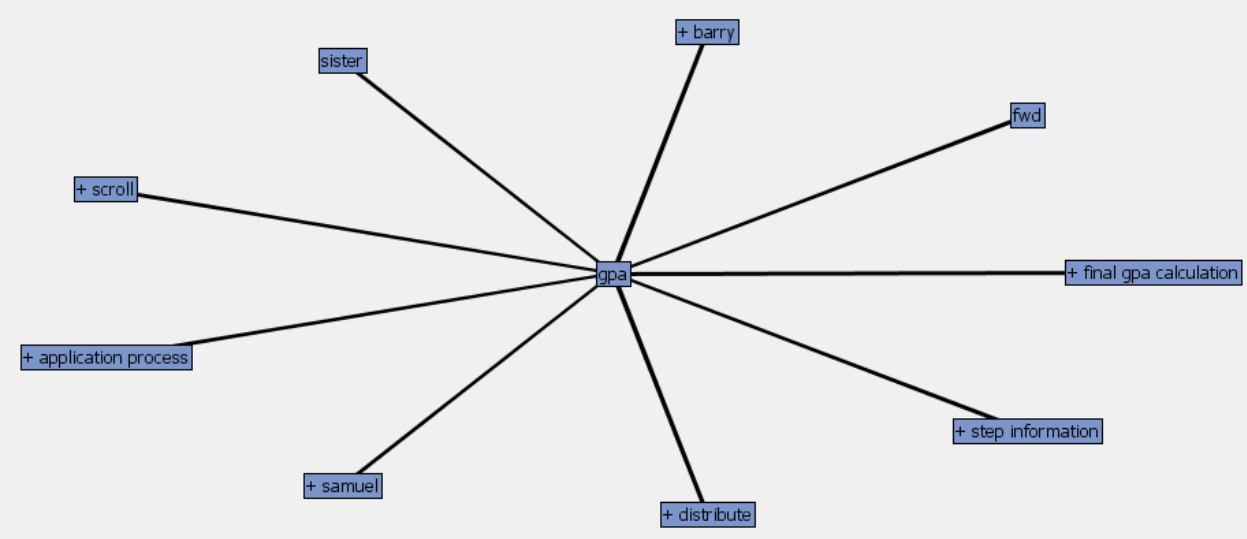
This cluster is concerned with students who want to take a transfer to OSU. Flower mound might represent the place in Texas from where a large number of students enroll into OSU.

Based on this, the university can conduct outreach programs in Flower Mound to increase conversion of those transfer students.

The words statistics and north might appear since OSU's North Classroom Building houses courses in Statistics and this might not represent anything conclusive for our analysis.

Topic 6: weston, sat, gpa, academic, high

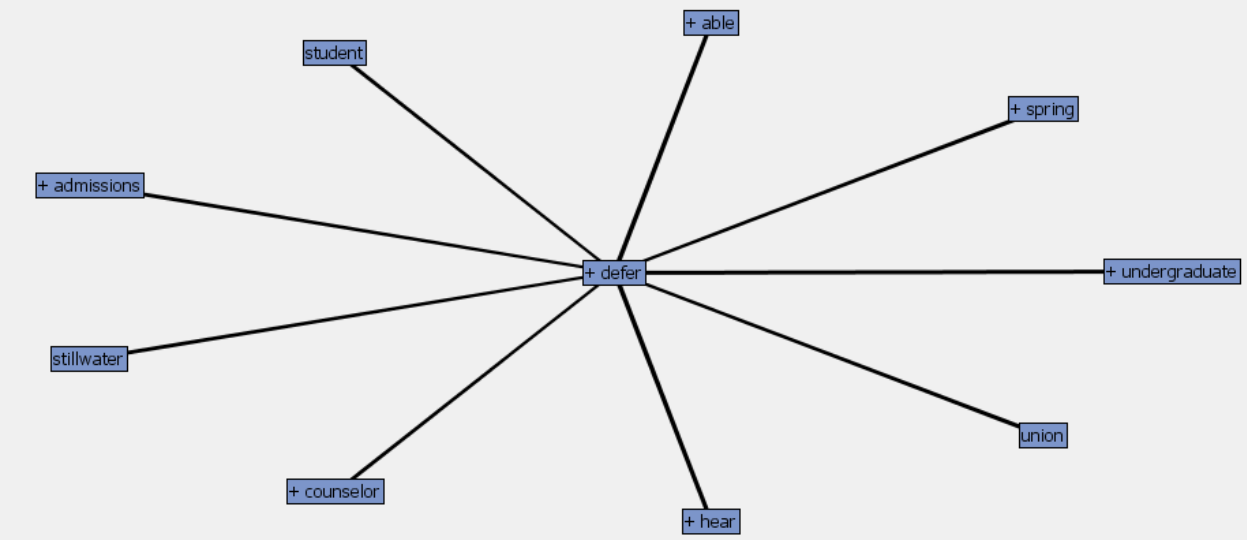
The student here is asking questions about gpa like how to calculate gpa or what high school gpa is required to get an admit into OSU. The university can clearly mention the gpa calculation and requirements on its website to address these concerns.



The concept link diagram above shows that the term is strongly linked with final gpa calculation, step information and application process which are most probably the concerns that students face regarding gpa.

Topic 7: defer, whaley, madi, cancel, register

Here the student might be asking about deferring their admission process. They might have gotten an admit but may not have plans to join immediately. They might be curious to know the process of joining the university on a later semester.



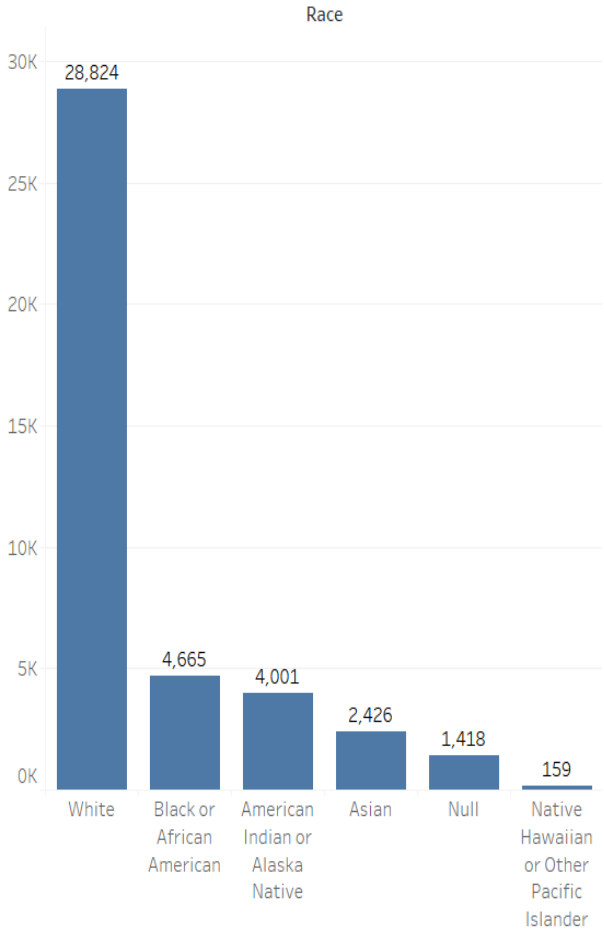
The concept link diagram for the word defer also shows that its strongly linked with terms like admissions, spring, counselor, undergraduate which confirms the above point.

When such a situation arises where the student wants to defer the admission, the university should try to find the reason for it and address the issue or maybe provide some kind of financial aid as an incentive for the student to join the same semester.

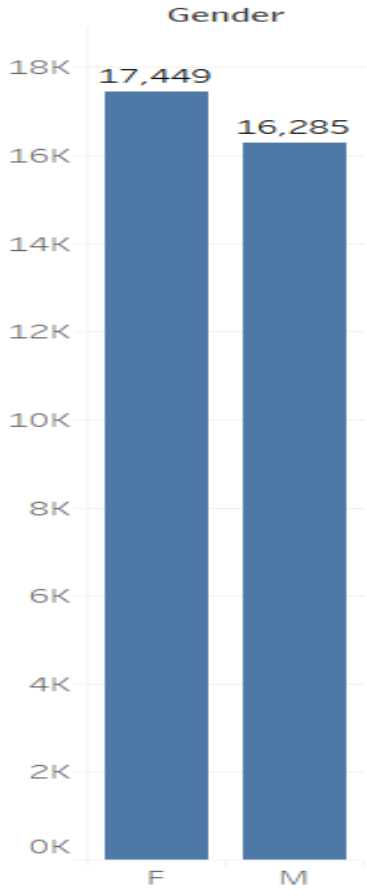
Descriptive Statistics:

The descriptive statistics of the students who were given and accepted an offer is shown below.

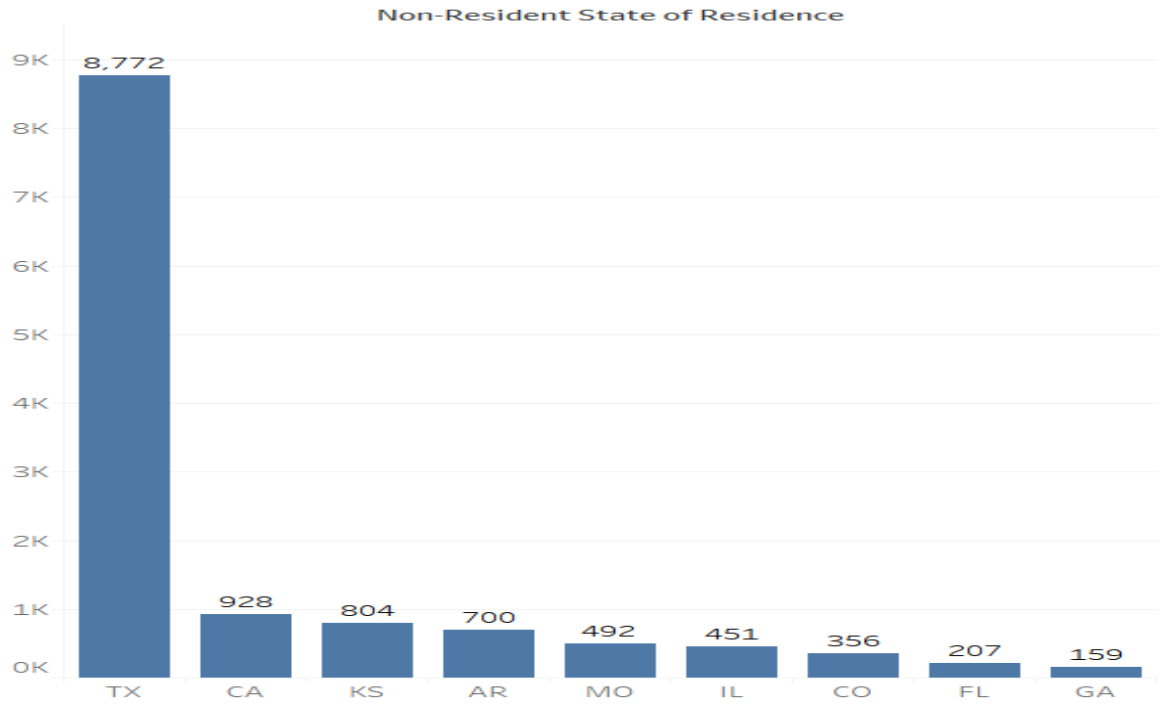
By Race:



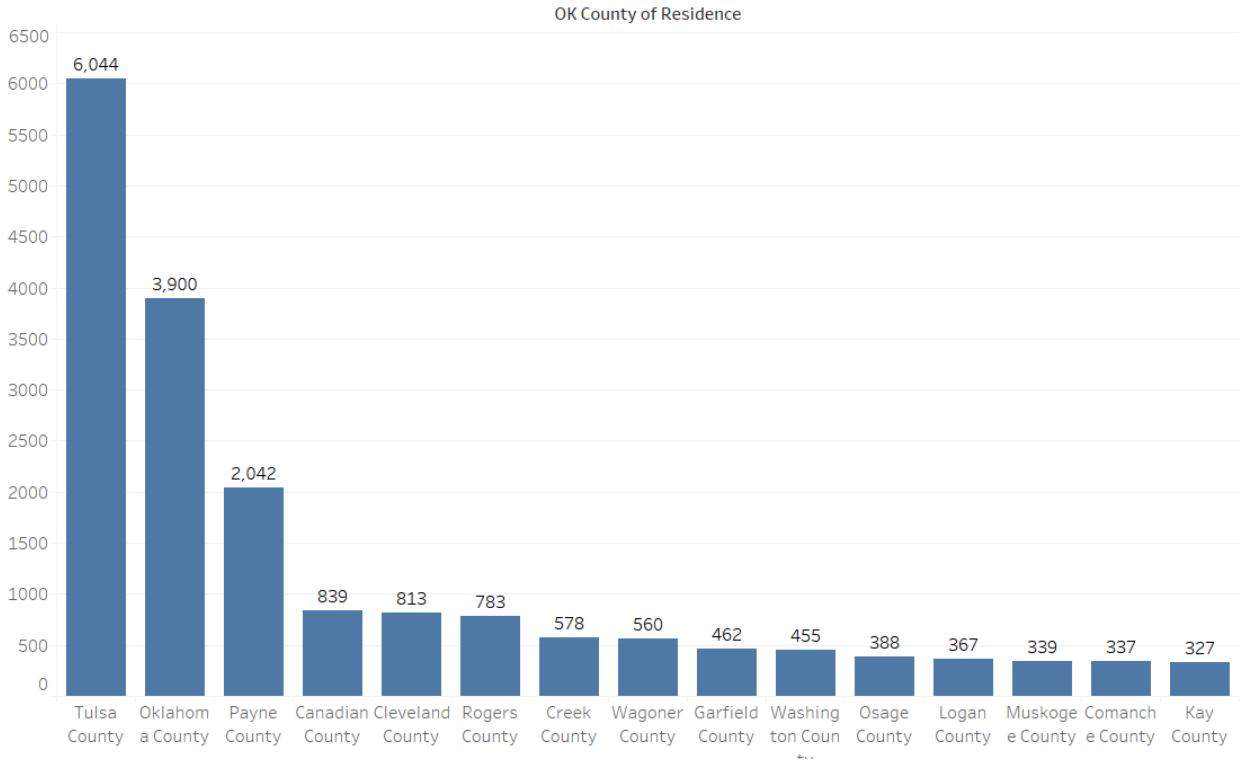
By Gender:



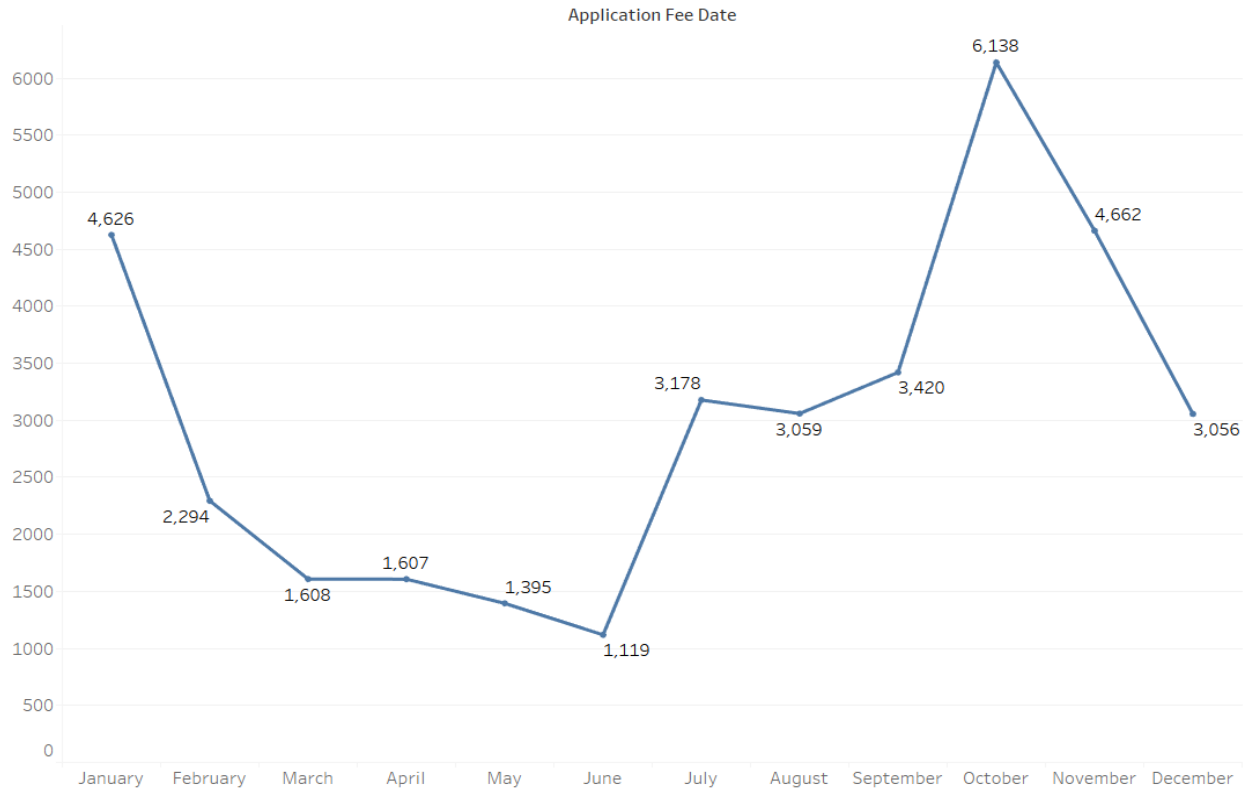
By Top Non Resident State of Residence:



By Top OK County of Residence:



By Month of Application Fee Date:



Conclusions

From the analysis, the students more likely to accept an admit and enroll into a university are the ones who enquire about terms like financial aid, counselor, transfer/defer options and academic gpa. It was also identified that an average of 4.6 student initiated emails are exchanged between the student and the university. Any number below 2 shows that the student will most likely not join the university.

A research on public school students in Boston area suggests that with proactive outreach to the student about the admission process including fee calculation and gpa requirements among other factors, the enrollment increased by 12%. It also suggests that the students struggle to decode 'highly confusing' financial aid letters and other barriers which often leads to the student putting off key tasks. This situation can be avoided if the university understand better what kind of issues each student faces and proactively answer those questions.

From the demographics, we can infer that most students from out of state belong to Texas followed by California and Kansas. The gender distribution is almost equal with Females leading marginally. Tulsa County, Oklahoma County ad Payne County are the top 3 counties from where the maximum students accept admission offers and October is the month when most students pay their application fees which is justified since the college application window starts from around the same time.

References

Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, *2002*(113), 17-36.

https://www.vox.com/2018/8/3/17639142/poor-kids-college-dont-enroll?linkId=55392932

https://support.sas.com/documentation/onlinedoc/guide/tut42/en/menu.htm

Bennett, J. and Lanning, S. 2007. The Netflix Prize. Proceedings of KDD Cup and Workshop 2007, San Jose, CA. Aug 12, 2007.

Huayi Li, Arjun Mukherjee, Jianfeng Si and Bing Liu. Extracting Verb Expressions Implying Negative Opinions. Proceedings of Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15). 2015.

Learning Word Vectors for Sentiment Analysis by Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

Acknowledgements

I wish to express my sincere gratitude to Dr Goutam Chakraborty for his teachings and support during the course of this paper. I would also like to sincerely thank Dr Miriam McGaugh for her constant support and guidance throughout the project.

Contact Information

Shashikant Chebrolu

Graduate Student of Business Analytics

Oklahoma State University

shashi.chebrolu@okstate.edu

Phone: 405 612 3883

LinkedIn: https://www.linkedin.com/in/shashichebrolu/