Salary Prediction

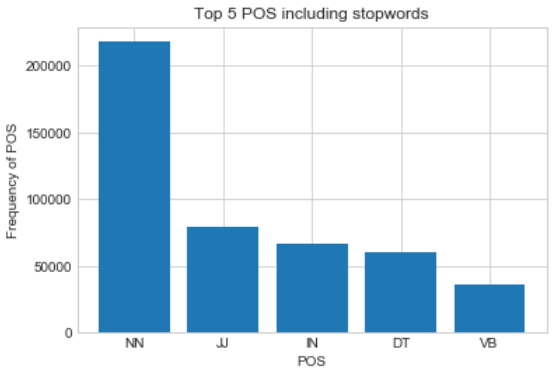
First, we remove rows which contain NA and randomly sampled 2500 data from clean dataset. After removing punctuation, converting words to lower cases, and performing lemmatization, the top 5 parts of speech (POS) in the job descriptions are:

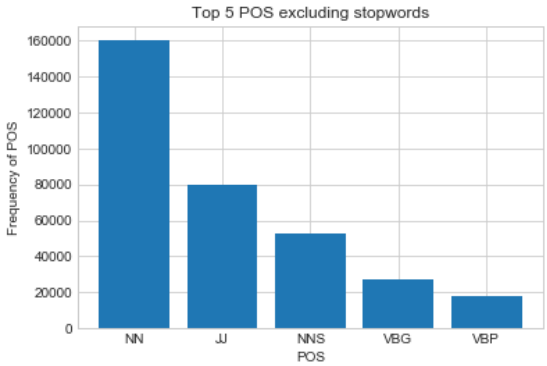
Rank	Frequency	POS	
1	218091	NN	Noun, singular or mass
2	78915	JJ	Adjectives
3	66451	IN	Preposition or subordinating conjunctions
4	60329	DT	Determiner
5	36068	VB	Verb, base form

When we exclude the stopwords, the top 5 parts of speech in the job descriptions become:

Rank	Frequency	POS	
1	159745	NN	Noun, singular or mass
2	79563	JJ	Adjectives
3	52912	NNS	Noun, plural
4	26952	VBG	Verb, gerund or present participle
5	17679	VBP	Verb, non-3rd person singular present

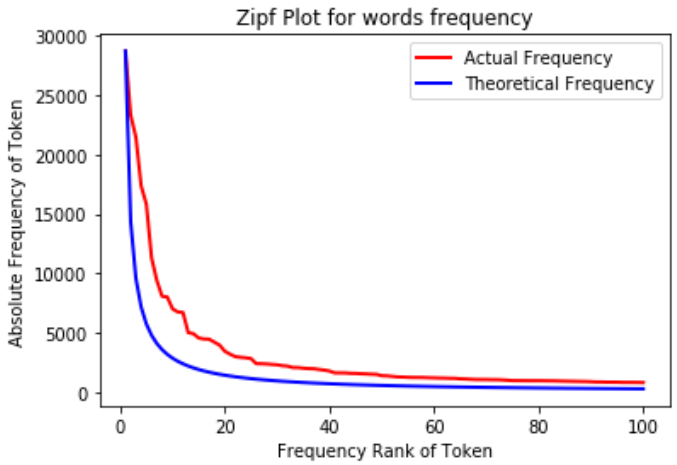
Below are the histograms for the top 5 parts of speech under two stated scenarios for better visualized comparison.





Regardless of stopwords, nouns and adjectives are the two most occurring parts of speech for all job descriptions in the data. This is comprehensible since people need to understand specific positions, generally in noun form, with its characteristics, often in adjectives, describing the positions. However, without stopwords, the top third becomes plural nouns. Stopwords are mostly likely prepositions, subordinating conjunctions, such words can be “before, after, until, etc.”, possibly indicating the time-related aspect of the job postings, determiners such as “the, a, this, etc.”, and verbs in base form for this dataset.

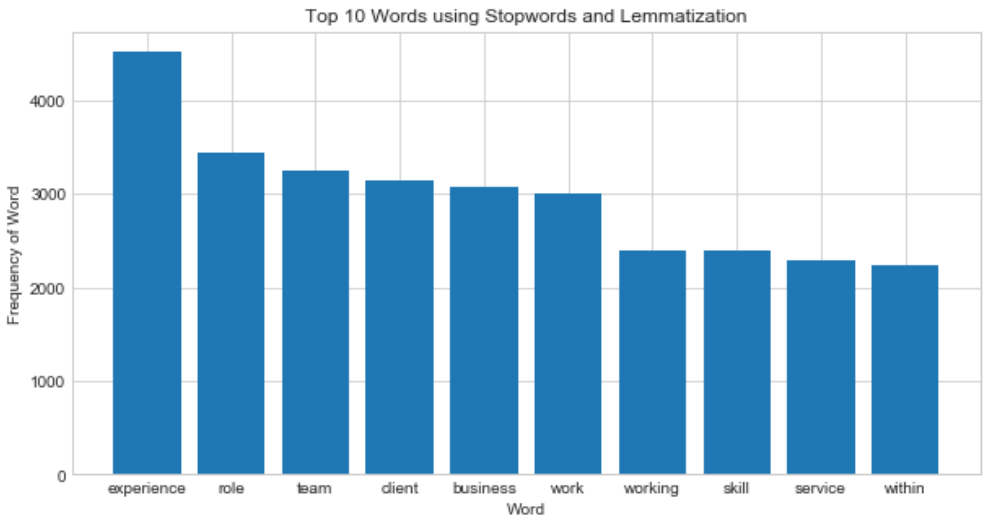
Without performing stemming or lemmatization, and keeping the stopwords, we plot the rank instead of actual words to prevent unreadable axis. The graph for most common 100 words in the data against the theoretical prediction of Zipf’s law is:



This data supports Zipf’s Law. According to Zipf’s Law, frequency of a word is inversely proportional to its rank. As can be seen from the graph above, the trend of actual frequency follows closely with the theoretical frequency of the law. There are few words that occur very frequently, a medium number of words have moderate frequency, and many words occurs very infrequently.

After removing stopwords and lemmatizing the data, the 10 most common words are:

Rank	Word	Frequency
1	experience	4511
2	role	3440
3	team	3239
4	client	3143
5	business	3078
6	work	3006
7	working	2401
8	skill	2388
9	service	2281
10	within	2234



With text mining, it seems like candidates with experience, who have worked in team, have the skills, providing services, possibly working with clients are the ones that recruiters are most commonly looking for in the job description.

Modelling

Model based on numeric predictors

Before starting building models, we make some assumption regarding missing value in the columns since there are a lot of missing value in contract time and contract type. For contract type, we assume that if it is not part-time, it should be full-time since most available job openings are generally looking for full-time positions. For contract time, we assume if it is not contract, it should be permanent for the same reasoning. We decide to use the Binominal model in the Naïve Bayers classifier because we only have two categories.

By searching online, we found that the following cities have the highest cost of living, which are London, Oxford, Brighton, Hove, Cambridge, Bristol, Reading, Berkshire, York, Portsmouth, Edinburgh, and Leeds. By using a Binominal model, we obtained an accuracy of 75.60%, which is calculated by using the total number of correct predictions divided by the sample size. This score means that how accurate the model is by matching predicting and the actual result for both high and low salary. The model's precision is 70.00%, which is calculated by using the accurate high salary prediction divided by the total number of high salary prediction. The model's recall is 10.77%, which is calculated by using the correct high salary prediction divided by the both correct high and low salary prediction. The confusion matrix is as followed.

	Predicted Low	Predicted High
Actual Low	364	6
Actual High	116	14

Model based on Text-based Predictors

By using a Binominal model, we obtained an accuracy of 77.20%, which is calculated by using the total number of correct predictions divided by the sample size. This score means that how accurate the model is by matching predicting and the actual result for both high and low salary. The model's precision is 59.30%, which is calculated by using the accurate high salary prediction divided by the total number of high salary prediction. The model's recall is 39.23%, which is calculated by using the correct high salary prediction divided by the both correct high and low salary prediction. The confusion matrix is as followed.

	Predicted Low	Predicted High
Actual Low	335	35
Actual High	79	51

The following table shows the top 10 words for both high and low salary.

Rank	High Salary	Low Salary
1	Business	School
2	Project	Experience
3	Experience	Role
4	Manager	Service
5	Team	Sale
6	Role	Client
7	Client	Teacher
8	Management	Team
9	Financial	Work
10	Finance	Manager

As we can see from the table, high salary job and low salary job have almost distinct keywords. We can also feel that high salary jobs have more management side key words whereas low salary jobs are more often related to school and sale.

Hybrid model—incorporating both numerical and text based predictors

By using a Binominal model, we obtained an accuracy of 76.60%, which is calculated by using the total number of correct predictions divided by the sample size. This score means that how accurate the model is by matching predicting and the actual result for both high and low salary. The model's precision is 57.47%, which is calculated by using the accurate high salary prediction divided by the total number of high salary prediction. The model's recall is 38.46%, which is calculated by using the correct high salary prediction divided by the both correct high and low salary prediction. The confusion matrix is as followed.

	Predicted Low	Predicted High
Actual Low	333	37
Actual High	80	50

Which model – numeric only, text only and hybrid – provided the highest accuracy in predicting high/low salary?

The text only model provided the highest accuracy in predicting high/low salary, which is an accuracy of 77.20%. The result surprised us because we thought the hybrid model, which combines both numeric and text, would deliver the best result because it considers more factors and the job features variables may help the model to predict more accurately. However, because of the abundance information, it may create too much noises into the model. The hybrid model cannot deliver the best result. This comparison also shows us text analytics is a powerful tool, and the job posting will use specific wording for different type of job. Therefore, even though some job may have the same feature, the wording will have more power in determining the job's salary information. Moreover, model from B1 generates lowest accuracy because the same variables we use in B1 can be applied to both low and high salary categories. Those variables seem to not have strong predictive power comparing to the text model.