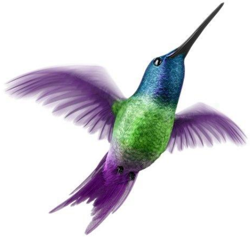
SOW Risk Scoring Model

Hafiza Umair, Khaled AlNajjar, Kiranpal Sohi, Ramona Torabian

April 7, 2020

TELUS

Introduction



Objective:

- Using NLP techniques, extract textual contents and clauses from SOW (Statement of Work) with TELUS suppliers/vendors
- Identify risk and label each SOW based on its risk level.
- Use unsupervised algorithm, namely clustering model to identify risk.

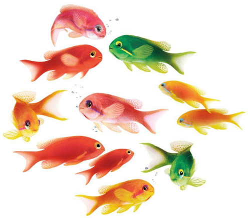
Method:

- Extract data from publicly available PDF SOWs downloaded from http://www.pdfsearchengine.net/
- Specific Data as required by Telus includes: 1) Work Description, 2) ETA or delivery date, 3) Headcount & rates, 4) Nature of resources, 5) Location of work, 6) Committed Funds, 7) Committed volume, 8) Currency (USD or CAD), 9) SOW Category (IT, fleet, field services, consulting, cleaning).
- Use extracted data to cluster the individual SOWs as High, Medium or Low risk.

Limitations:

- Documents were not of high quality as they were downloaded from internet. The content and subject of SOWs varied significantly depending on what was available freely online.

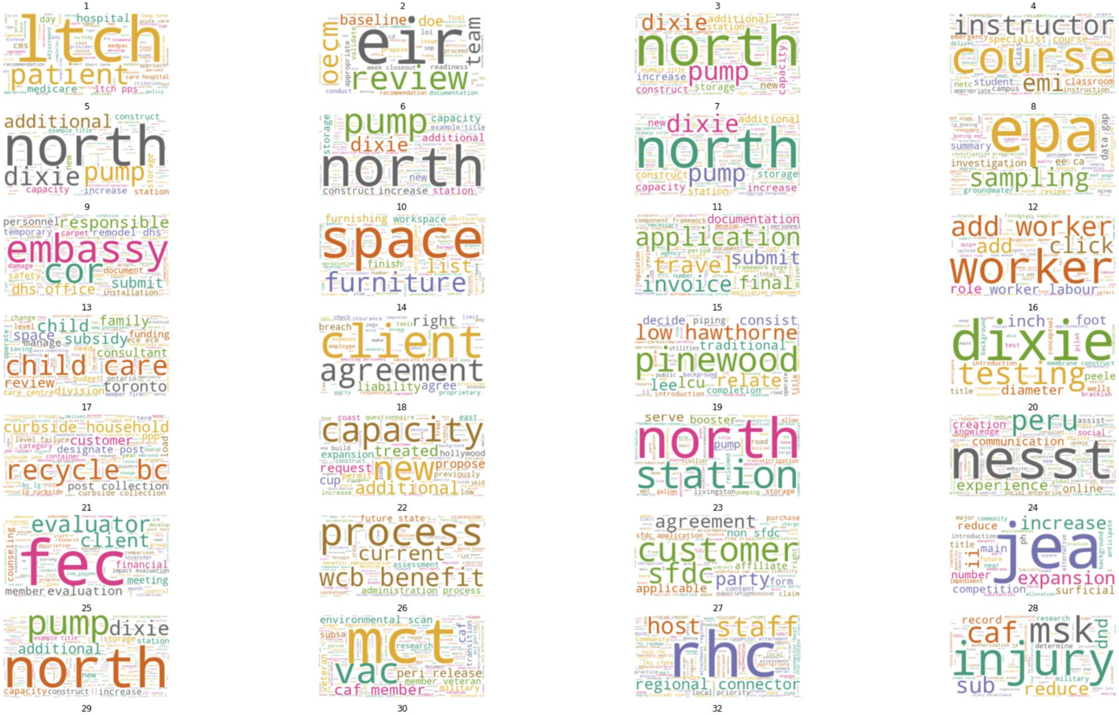
Preparation



- A total of 48 PDF Statement of Works were downloaded from the Internet.
- The content varied from software to hospital system solution to Construction.
- A total of 30 PDFs were readable and uploaded to Google Colab notebook for code development.
- Used NLTK, and spaCy:
  - Clean up data - removing stopwords, lower case, punctuation removal
  - Lemmatization applied
- Built Data Corpus - bag of words
- Built Document Term Matrix - DTM using CountVectorizer
  - Numerical representation of the text so that machine can read and process
- Analyzed common words in all SOWs, added more stopwords and rebuilt DTM

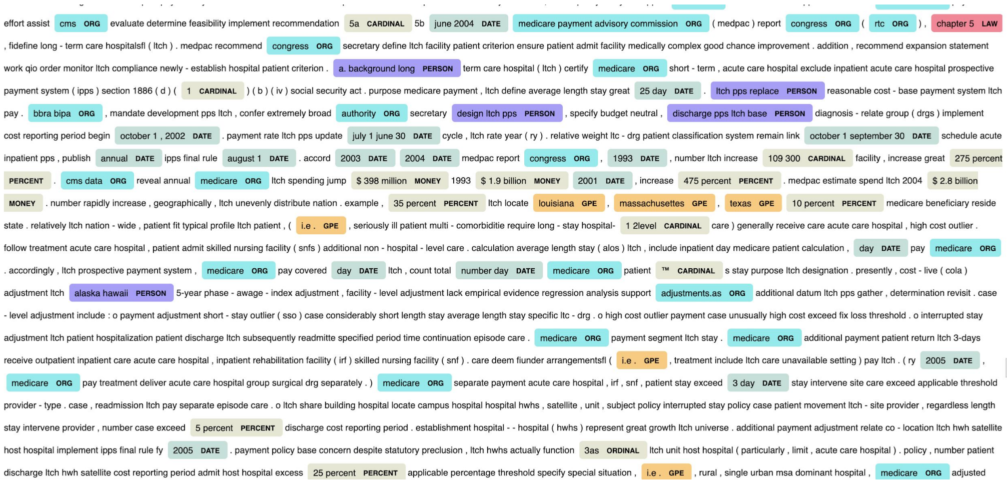
Word Cloud Analysis

- Using spaCy library, the Word Clouds were created for analysis.
- Prominent words (high frequency) in Word Cloud provide quick understanding of each individual PDF SOW.
- The subject of SOWs varied from including: medical, furniture, education, technology.



spaCy Word Label

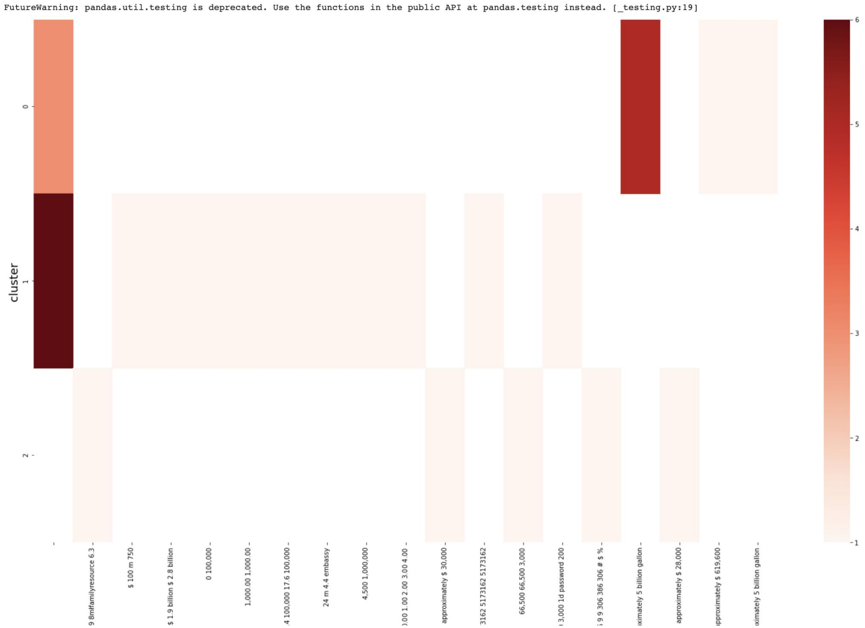
- Using spaCy library, standard word labels were used to identify the following:
  - ETA or Delivery Date
  - Committed Funds
  - Location of Work
  - Committed Volume
- For the following key items the code is still under development:
  - Work Description
  - Headcount & Rates
  - Nature of Resources
  - Currency



effort assist cms ORG evaluate determine feasibility implement recommendation 5a CARDINAL 5b june 2004 DATE medicare payment advisory commission ORG (medpac) report congress ORG (rtc ORG) , chapter 5 LAW
, fidefine long - term care hospitalsfl (ltch) . medpac recommend congress ORG secretary define ltch facility patient criterion ensure patient admit facility medically complex good chance improvement . addition , recommend expansion statement
work qio order monitor ltch compliance newly - establish hospital patient criterion . a. background long PERSON term care hospital (ltch) certify medicare ORG short - term , acute care hospital exclude inpatient acute care hospital prospective
payment system (ipps) section 1886 (d) (1 CARDINAL) (b) (iv) social security act . purpose medicare payment , ltch define average length stay great 25 day DATE . ltch pps replace PERSON reasonable cost - base payment system ltch
pay . bbra bipa ORG , mandate development pps ltch , confer extremely broad authority ORG secretary design ltch pps PERSON , specify budget neutral , discharge pps ltch base PERSON diagnosis - relate group (drgs) implement
cost reporting period begin october 1 , 2002 DATE . payment rate ltch pps update july 1 june 30 DATE cycle , ltch rate year (ry) . relative weight ltc - drg patient classification system remain link october 1 september 30 DATE schedule acute
inpatient pps , publish annual DATE ipps final rule august 1 DATE . accord 2003 DATE 2004 DATE medpac report congress ORG , 1993 DATE , number ltch increase 109 300 CARDINAL facility , increase great 275 percent
PERCENT . cms data ORG reveal annual medicare ORG ltch spending jump $ 398 million MONEY 1993 $ 1.9 billion MONEY 2001 DATE , increase 475 percent PERCENT . medpac estimate spend ltch 2004 $ 2.8 billion
MONEY . number rapidly increase , geographically , ltch unevenly distribute nation . example , 35 percent PERCENT ltch locate louisiana GPE , massachusettes GPE , texas GPE 10 percent PERCENT medicare beneficiary reside
state . relatively ltch nation - wide , patient fit typical profile ltch patient , (i.e . GPE , seriously ill patient multi - comorbiditie require long - stay hospital- 1 2level CARDINAL care) generally receive care acute care hospital , high cost outlier .
follow treatment acute care hospital , patient admit skilled nursing facility (snfs) additional non - hospital - level care . calculation average length stay (alos) ltch , include inpatient day medicare patient calculation , day DATE pay medicare ORG
. accordingly , ltch prospective payment system , medicare ORG pay covered day DATE ltch , count total number day DATE medicare ORG patient ™ CARDINAL s stay purpose ltch designation . presently , cost - live (cola)
adjustment ltch alaska hawaii PERSON 5-year phase - awage - index adjustment , facility - level adjustment lack empirical evidence regression analysis support adjustments.as ORG additional datum ltch pps gather , determination revisit . case
- level adjustment include : o payment adjustment short - stay outlier (sso) case considerably short length stay average length stay specific ltc - drg . o high cost outlier payment case unusually high cost exceed fix loss threshold . o interrupted stay
adjustment ltch patient hospitalization patient discharge ltch subsequently readmitte specified period time continuation episode care . medicare ORG payment segment ltch stay . medicare ORG additional payment patient return ltch 3-days
receive outpatient inpatient care acute care hospital , inpatient rehabilitation facility (irf) skilled nursing facility (snf) . care deem fiunder arrangementsfl (i.e . GPE , treatment include ltch care unavailable setting) pay ltch . (ry 2005 DATE ,
medicare ORG pay treatment deliver acute care hospital group surgical drg separately .) medicare ORG separate payment acute care hospital , irf , snf , patient stay exceed 3 day DATE stay intervene site care exceed applicable threshold
provider - type . case , readmission ltch pay separate episode care . o ltch share building hospital locate campus hospital hospital hwhs , satellite , unit , subject policy interrupted stay policy case patient movement ltch - site provider , regardless length
stay intervene provider , number case exceed 5 percent PERCENT discharge cost reporting period . establishment hospital - - hospital (hwhs) represent great growth ltch universe . additional payment adjustment relate co - location ltch hwh satellite
host hospital implement ipps final rule fy 2005 DATE . payment policy base concern despite statutory preclusion , ltch hwhs actually function 3as ORDINAL ltch unit host hospital (particularly , limit , acute care hospital) . policy , number patient
discharge ltch hwh satellite cost reporting period admit host hospital excess 25 percent PERCENT applicable percentage threshold specify special situation , i.e . GPE , rural , single urban msa dominant hospital , medicare ORG adjusted

K-Mean Cluster

- The text content of PDFs was used to develop a clustering algorithm.
- The high risk SOWs based on Committed Funds ($s) is Cluster 0: Construction Industry based SOWs
- The value of Medium and Low risk SOWs were difficult to assess as the value of Committed Funds were either not in the publicly available SOWs or were difficult to extract for analysis.

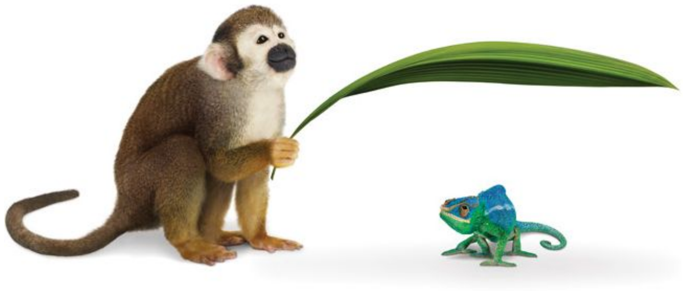


```
FutureWarning: pandas.util.testing is deprecated. Use the functions in the public API at pandas.testing instead. [_testing.py:19]
```

Clusters

```
 0 : water, aquifer, floridan, well, reclaim, mgd, phase, supply, wellfield, construction
 1 : contractor, shall, task, work, service, eir, client, project, ltch, sfdc
 2 : child, caf, vgisc, injury, care, mct, city, wcb, process, service
```

Conclusion



- Word Cloud provided a visual analysis of each individual PDF Statement of Work. It can be used by Telus to quickly determine the keywords and subject matter of SOW.
- spaCy library provided a good means to clean up the data and allowed for extraction of approximately half of the key data from SOWs. This process can be used to extract data from Telus SOWs.
- K-Mean Clustering can be used to group SOWs into High, Medium and Low risk. With Telus, the volume of PDFs used for clustering can help enrich the clustering learning and provide a more robust clusters analysis.

Improvement Opportunities

- Identifying the confidence intervals
- Using Custom Name Entity Recognition - NER to train spaCy model and better clean the data for further analysis and increase accuracy.
- K-Mean Cluster analysis requires cleaner and greater volume of data to ensure robust results.
- Additional PDFs are required and if possible with similar structure. Telus PDFs may be more consistent and therefore easier to extract data from.