

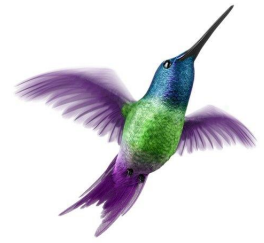
SOW Risk Scoring Model

Hafiza Umair, Khaled AlNajjar, Kiran Sohi, Ramona Torabian

April 7, 2020



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.

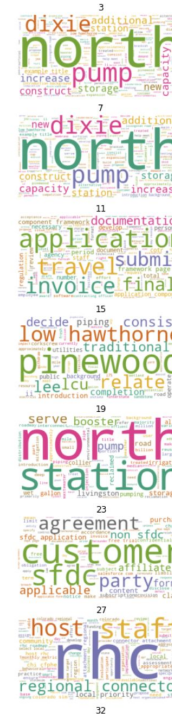
Preparation



- 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:

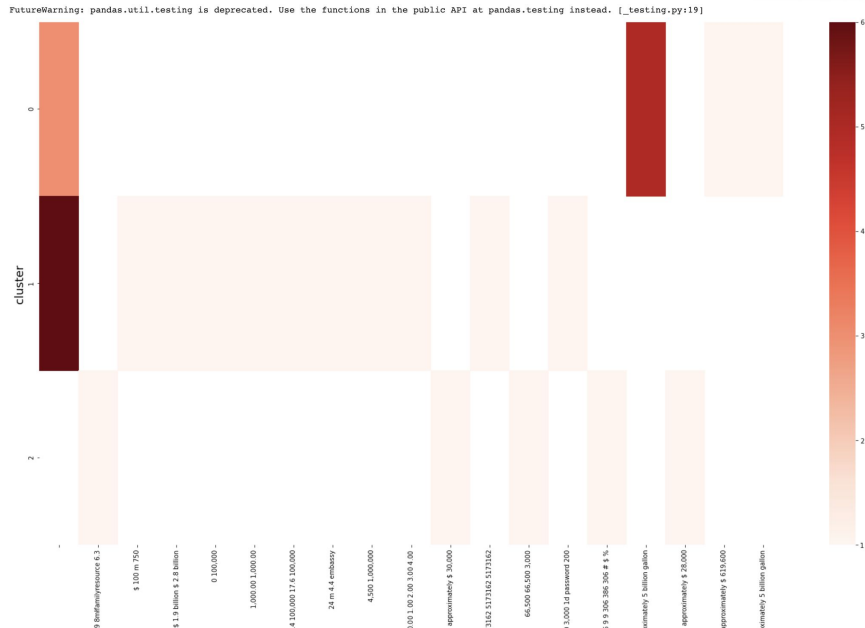
- Work Description
- Headcount & Rates
- Nature of Resources
- Currency

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

K-Mean Cluster



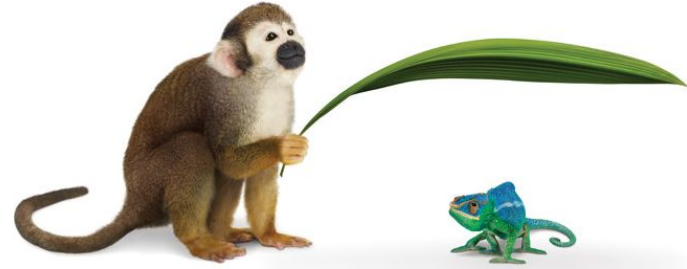
- The text content of PDFs was used to develop a clustering algorithm.
- The high risk SOWs based on Committed Funds (\$\$) is Cluster 0: Construction Industry based SOWs
- The value of Medium and Low risk SOWs were difficult to assess as the value of Committed Funds were either not in the publicly available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



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.

