Theory Qns:

1. Problem 1-1 Text Analytics Pipeline

Qns1: Within the last year, an insurance company has transitioned from physical printouts to digitized docu- ment scans for all customer-related forms. These documents should now be used for further analysis of customer needs.

Briefly describe each of the planning considerations for a text analytics project in the above scenario. As a reminder, the considerations include:

(i) drivers, (ii) objectives,

(iii) data availability, and (iv) cost factors

1.1.1

- Driver: since the main oncern is still to be decided based on the data obtained, this would be the typical text-driven problem where we derive conclusions and decisions stemming from the collected data itself.
- Objective: the main objective of analyzing costumer data is to reach higher profit by better tailoring to specific customer needs.
- Data Availability: The scanned version of insurance paperworks would need further processing: conversion into raw text, cleaning, parsing etc.
- Cost factors: The cost of data gathering must be considered as necessary conversion&cleaning should be carried out, as well as the possible tranformation between formats.

Qns2: Compare the usage of Elasticsearch as a document store over a traditional relational database, such as PostgreSQL. Give two advantages and disadvantages.

1.1.2

- Advantage:

Elasticsearch adapts the inverted index system of Lucene to offer speedy full-text searches, while traditional RDBMS based on SQL query could be relatively slower in response. The structuralized JSON documents where elasticsearch stores its data could be visited from any node instantly, reducing the number of read operations required.

Elasticsearch does not require user to specify the schema of its index; it performs well in auto-inference for datatypes such as number, boolean values and timestamps, offering certain flexibility in storage, while PostgreSQL operates on a rather rigid schema definition basis.

- Disadvantage:

The above can be also the drawback for elasticsearch, as PostgreSQL offers a plethora of functionalities based on its support of multiple data types.

Elasticsearch also sacrifices certain functionalities---such as transaction---for higher speed, while PostgreSQL supports a broader range of mechanisms including a robust transaction mechanism, while user can either roll back the operation to a given point or bundle operation together.

Elasticsearch has a more loose security system where every user is granted the same access as admin, while PostgreSQL tend to apply strict access control with multiple authentications for example LDAP etc.

Generally, based on the CAP theorem, what Elasticsearch offers is Availability and Partition Tolerance, while PostgreSQL offers Consistency and Availability.

Qns3: Name at least two advantages and disadvantages for each of the following concepts. Illustrate your points by giving concrete examples.

(i) Stop word removal (ii) Stemming

1.1.3 Stop word removal

- Advantage:

1. In certain scenarios where stop words are trivial (sentiment analysis, theme classification etc.), its removal would reduce corpus size greatly and keep the information provided by the corpus clean-cut and better fitting for classification or clustering. Example:

"There are 5 people who are unsatisfied about the Mensa food today."->["unsatisfied" "food"].

The important keywords for sentiment are preserved while trivialities are cut down, word count decrease from 12 to 2. It lowers both the space dimension and data size greatly. 2. As the more significant words are preserved, the enhanced datasets with less noise should benefit training accuracy, training speed and model performance.

- Disadvantage:

1. Inappropriate stop word removal could change the meaning of text drastically. For example if consider "not" as a stop word (as in Aruana) and remove it, "The food is not good" -> ["food" "good"], which is misleading. Sentiment analysis is sensitive to the stop words one select, and one should be careful of not leaving deciding features out.

2. Object consisting of certain "stop words" would be unintelligible. example for search engine:

"to be or not to be, that is the question"->["question"]

"take that" ->[]

3. Problems arise because of loss of contexts due to stop word removal.\

Example as on slides:

["Jill" "reported" "CEO" "company"]<-"Jill reported to the CEO of the company" or "Jill reported as the CEO of the company".

If the model is not word-based but sequence-based, the training would encounter problems due to linguistic ambiguity. Thus stop word removal should not be considered if the preprocess is not specifically for word-based methods.

Stemming

- Advantage:

1. reduces size of datasets when using bag-of-words methods for assembling different forms of the same word together. For example:

["liking" "likes" "liked"] -> ["like"]

2. increases processing speed. Stemming, as the purely algorithmic process, is often faster to use than applying word variants, since it does not require any further operation or context information.

- Disadvantage:

1. Stemming certain words is not possible, as stemmers regularly can't recognize variants that deviate from the general grammar rules for example "good" and "better" and "best". ("So why not substitute it with "goodest"?"...Here is when the Orwellian Newspeak comes into use.)
2. Stemmed words can be ambiguous, for example Axes is plural for both Ax and Axis. This is called under-stemming, as roots and words are not always corresponding on a one-to-one basis.
3. There is also the phenomenon of over-stemming when words of very different meaning are stemmed into the same root. Basic example: while "the likes of" means "things similar to", "like" is also used as a verb. When "likes" as noun is stemmed and you are doing word-based sentiment analysis, it might be misleading. More example: "farther" and "farthing" would be both stemmed into "farth" while their meanings differ greatly.

Qns2:Why is a high quality conversion from PDF to plain text hard? Your answer does not need to be exhaustive but should outline some of the most important reasons.

A PDF page is drawn based on instructions contained in the content stream. Following are the procedures which are difficult to compute while converting from pdf to plain text:

1. It is also possible to draw text in areas where text extractors do not look: some tools hide text by placing it in a pattern and filling the page area with that pattern; similarly, type 3 fonts are available; each character in a type 3 font has its own content stream; therefore, a tool can draw all the text within a single glyph of a type 3 font and then draw it on the page.

2. The process of reading through PDFs line by line may seem straightforward. There can be differences in the way documents are structured, and data can be arranged ad hoc. Also, documents can be skewed in various ways, making them difficult to read.
3. There are many PDFs that follow different document structures, so the data front, the size, and the angle can differ from one PDF to another. Therefore, it will be more difficult to process data is challenging.

Qns 4: Apply your solution from task4 (i)tothefilesinthefolderscans/whichconsistsofpagesscannedfrom a phone book. Analyse how well your solution performs by giving examples.

This mupypdf method as considered scanning this pdf column-vise, Based on ads they are 4 columns in the original scanned OCR.

Consider an ad as shown below:

After using Mupypdf, below is the text format conversion:

```
'Alt Peter \n',
 'Rechtsanwalt \n',
 'Fachanwalt für Arbeitsrecht \n',
 'Kirschgartenstr. 19 \n',
```

```
 '33836 70 \n',
 'www.kanzlei-buehler-alt.de \n',
 'm) \n',
 'Gratis anrufen | mit dastelefonbuch.de \n',
 '  \n',
 'Bau, Becky & Collegen \n',
 'Poststr. 2 \n',
 'www.bau-coll.de \n',
 'ed al Gratis anrufen \n',
 '| mit dastelefonbuch.de \n',
```

Most of the text from the sample has been detected by the method except in this case the phone from Bau,Becky & Collegen has not been detected.

Example2:

Here, the telephone symbol has not been detected properly

```
'Jakob & Kollegen \n',
 'Jakob & Kollegen \n',
 'Rechtsanwälte \n',
 'Bergheimer Straße 49 \n',
 '| 69115 Heidelberg \n',
 '2 (0 62 21) 14 07 14 \n',
 'www.jakob-ra.de \n',
```

Now considering the single OCR,

The method has run from left to right but not in columns as doubleOCR

```
'Altlußheim (0.62.05) \n',
 '| \n',
 'ne. \n',
 'www.dastelefonbuch.de \n',
 'Gottfried Werner Tulla-20 ........020...0.... 32030 | Heid Dagmar
Rosemarie.................. 364 1733 | Hoffstätter Dieter u. Fischer
Beate......... 31973 \n',
 '  \n',
 '  \n',
 '  \n',
```

```
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '     \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 '  \n',
 'AESE5-APOTHEKE| \n',
 'cottmam Eisabeth Hause \n',
 '31942 | \n',
 'Erich-Kästner-Weg 1A \n',
 'Kurpfalz-9 \n',
 'Gottwald G. Fliederweg 6 .................... 834139 | - Dagmar
Rosemarie............. 01791173263 | - Elisabeth Kurpfalz-9
..................... 34747 \n',
 'Tel. 062 21-983 13 31 \n',
 'Gottwald-Münch Uta Fliederweg 6 ......... 32648 | \n',
 'Lußhardt-5 \n',
 '| \n',
 'Hofmann Walter Rheinhäuser-S4 .......... 32253 \n',
 '. \n',
 'ei \n',
 'Grabowski Wäscherei Schul- 2: aan 34341 | - Kurtund Dagmar... \n',
 '33023 | Hohn Elisabeth Haupt-53 ................... 34975 \n',
 'www.atos-apotheke.de \n',
 'Graf Madlen Haupt-9 .........u....... 953159 | \n',
 'Erich-Kästner-Weg 1a \n',
 '— Josef Hölderlin-9......................... 322 60 \n',
```

```
 'Grafe Rolf Kurpfalz-S1 ..... 32499 | Heidemann Rainer
u. Alexandra..... 32304 | - Robert Hölderlin-9..... 0178895
72 64 \n',
 'Apotheker M. Schuol \n',
```

Here the spelling from the ad it is AtoS APOTHEKE but it has read as AESE5 APOTHEKE.Its not much readable as the previous double OCR.