Content-based Frailty Classification System ++

Hospital Project

Project type: R&D (Competitive Grant)

Week 12 Final Presentation

Done By: Lee Jie Ying, Lynn (161015H), Low Ren Ern (163027J)

Overview of Presentation

- Project Scope
- Clinical Frailty Scale
- Handling 2018 Old and New Sample Data from TTSH
- Difference between Old and New Sample Data
- Timeline (Week 1-5)
- Timeline (Week 6-11)
- Task Accomplished
- Flowchart of Project
- Problems Encountered
- Solutions
- Summing-up

Project Scope

- This project is awarded under NYP-TTSH grant (18 Months)
- The purpose of this project is to develop a system to categorize the patient's notes to different frailty score.
- This is to improve the consistency of the score assignment so that they can render the suitable healthcare to the patient with the category.
- TTSH have provided 2015 and 2018 sample data that have been categorized for us to train the data.
    - Focus mainly on 2018 sample data

Clinical Frailty Scale (CFS)

- Objective: To examine the CFS in patients admitted to acute medical ward and its association with length of stay.
- Frailty Score (also known as category in this project)

Clinical Frailty Scale (CFS)

For example: Category 1 is extremely fit and Category 9 is terminally ill

Clinical Frailty Scale*

1 Very Fit – People who are robust, active, energetic and motivated. These people commonly exercise regularly. They are among the fittest for their age.

2 Well – People who have **no active disease symptoms** but are less fit than category 1. Often, they exercise or are very **active occasionally**, e.g. seasonally.

3 Managing Well – People whose **medical problems are well controlled,** but are **not regularly active** beyond routine walking.

4 Vulnerable – While **not dependent** on others for daily help, often **symptoms limit activities.** A common complaint is being "slowed up", and/or being tired during the day.

5 Mildly Frail – These people often have **more evident slowing,** and need help in **high order IADLs** (finances, transportation, heavy housework, medica-tions). Typically, mild frailty progressively impairs shopping and walking outside alone, meal preparation and housework.

6 Moderately Frail – People need help with **all outside activities** and with **keeping house.** Inside, they often have problems with stairs and need **help with bathing** and might need minimal assistance (cuing, standby) with dressing.

7 Severely Frail – **Completely dependent for personal care**, from whatever cause (physical or cognitive). Even so, they seem stable and not at high risk of dying (within ~ 6 months).

8 Very Severely Frail – Completely dependent, approaching the end of life. Typically, they could not recover even from a minor illness.

9 Terminally Ill - Approaching the end of life. This category applies to people with **a life expectancy <6 months**, who are **not otherwise evidently frail.**

Scoring frailty in people with dementia

The degree of frailty corresponds to the degree of dementia. Common **symptoms in mild dementia** include forgetting the details of a recent event, though still remembering the event itself, repeating the same question/story and social withdrawal.

In **moderate dementia**, recent memory is very impaired, even though they seemingly can remember their past life events well. They can do personal care with prompting.

In **severe dementia**, they cannot do personal care without help.

* 1. Canadian Study on Health & Aging, Revised 2008.
2. K. Rockwood et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005;173:489-495.

Handling 2018 Old Sample Data from TTSH

ED (Emergency) Notes Sample Data

Total: **44** files

- 860246760_Cat_5_ED Notes
- 861236598_Cat_4_ED Notes
- 862832645_Cat_6_ED Notes
- 866043327_Cat_7_ED Notes
- 866972291_Cat_6_ED Notes
- 868528977_Cat_5_ED Notes
- 869240003_Cat_7_ED Notes
- 872877037_Cat_6_ED Notes
- 883716700_Cat_6_ED Notes
- 860305644_Cat_5_ED Notes
- 861341492_Cat_5_ED Notes
- 864605611_Cat_6_ED Notes
- 866197419_Cat_3-4_ED Notes
- 867167556_Cat_4_ED Notes
- 868533852_Cat_5_ED Notes
- 869643579_Cat_7_ED Notes
- 882668111_Cat_8_ED Notes
- 883918817_Cat_4_ED Notes
- 860389965_Cat_5_ED Notes
- 861541971_Cat_5_ED Notes
- 864704750_Cat_6_ED Notes
- 866312687_Cat_3_ED Notes
- 867359638_Cat_6_ED Notes
- 868589632_Cat_6-7_ED Notes
- 869797158_Cat_8_ED Notes
- 882695046_Cat_7_ED Notes
- 889029982_Cat_6_ED Notes
- 860512095_Cat_5_ED Notes
- 861625560_Cat_7_ED Notes
- 864773682_Cat_7_ED Notes
- 866460588_Cat_6_ED Notes
- 867808171_Cat_4_ED Notes
- 868981160_Cat_6_ED Notes
- 872566043_Cat_5_ED Notes
- 882754748_Cat_4_ED Notes
- 892486974_Cat_7_ED Notes
- 860755305_Cat_3_ED Notes
- 861864723_Cat_6_ED Notes
- 864873792_Cat_7_ED Notes
- 866745356_Cat_7_ED Notes
- 868194296_Cat_5_ED Notes
- 869095984_Cat_6_ED Notes
- 872871395_Cat_5_ED Notes
- 883054456_Cat_7_ED Notes

IFA Sample Data

Total: **48** files

- 860246760_Cat_5_IFA
- 861236598_Cat_4_IFA
- 862832645_Cat_6_IFA
- 866197419_Cat_ 3-4_IFA
- 867167556_Cat_4_IFA
- 868528977_Cat_5_IFA
- 869095984_Cat_6_IFA
- 872871395_Cat_5_IFA
- 882695046_Cat_7_IFA
- 889029982_Cat_6_IFA
- 860305644_Cat_5_IFA
- 861341492_Cat_5_IFA
- 864605611_Cat_ 6_IFA
- 866312687_Cat_3_IFA
- 867359638_Cat_6_IFA
- 868533852_Cat_5_IFA
- 869240003_Cat_7_IFA
- 872877037_Cat_6_IFA
- 882754748_Cat_4_IFA
- 892486974_Cat_7_IFA
- 860389965_Cat_5_IFA
- 861541971_Cat_5_IFA
- 864704750_Cat_6_IFA
- 866460588_Cat_6_IFA
- 867773642_Cat_7_IFA
- 868589632_Cat_ 6-7_IFA
- 869643579_Cat_7_IFA
- 875352977_Cat_ 4-5_IFA
- 883054456_Cat_7_IFA
- 860512095_Cat_5_IFA
- 861625560_Cat_7_IFA
- 864773682_Cat_7_IFA
- 866745356_Cat_7_IFA
- 867808171_Cat_4_IFA
- 868981160_Cat_6_IFA
- 869797158_Cat_8_IFA
- 879996159_Cat_ 5-6_IFA
- 883716700_Cat_6_IFA
- 860755305_Cat_3_IFA
- 861864723_Cat_6_IFA
- 864873792_Cat_7_IFA
- 866972291_Cat_6_IFA
- 868194296_Cat_5_IFA
- 869060647_Cat_4_IFA
- 872566043_Cat_5_IFA
- 882668111_Cat_8_IFA
- 883918817_Cat_4_IFA

Handling 2018 Old Sample Data from TTSH

Patient Care Record Sample Data (consists of Admission & Discharge)

Total: **82** files

860305644_Cat_5_Patient Care Record (Inpatien...	860305644_Cat_5_Patient Care Record (Inpatien...	860389965_Cat_5_Patient Care Record (Inpatien...	860512095_Cat_5_Patient Care Record (Inpatien...	860755305_Cat_3_Patient Care Record (Inpatien...
860755305_Cat_3_Patient Care Record (Inpatien...	861236598_Cat_4_Patient Care Record (Inpatien...	861236598_Cat_4_Patient Care Record (Inpatien...	861341492_Cat_5_Patient Care Record (Inpatien...	861541971_Cat_5_Patient Care Record (Inpatien...
861541971_Cat_5_Patient Care Record (Inpatien...	861625560_Cat_7_Patient Care Record (Inpatien...	861625560_Cat_7_Patient Care Record (Inpatien...	861864723_Cat_6_Patient Care Record (Inpatien...	861864723_Cat_6_Patient Care Record (Inpatien...
862832645_Cat_6_Patient Care Record (Inpatien...	864605611_Cat_ 6_Patient Care Record (Inpatie...	864605611_Cat_ 6_Patient Care Record (Inpatie...	864704750_Cat_6_Patient Care Record (Inpatien...	864704750_Cat_6_Patient Care Record (Inpatien...
864773682_Cat_7_Patient Care Record (Inpatien...	864773682_Cat_7_Patient Care Record (Inpatien...	864873792_Cat_7_Patient Care Record (Inpatien...	866043327_Cat_7_Patient Care Record (Inpatien...	866043327_Cat_7_Patient Care Record (Inpatien...
866197419_Cat_ 3-4_Patient Care Record (Inpat...	866197419_Cat_ 3-4_Patient Care Record (Inpat...	866312687_Cat_3_Patient Care Record (Inpatien...	866312687_Cat_3_Patient Care Record (Inpatien...	866460588_Cat_6_Patient Care Record (Inpatien...
866460588_Cat_6_Patient Care Record (Inpatien...	866745356_Cat_7_Patient Care Record (Inpatien...	866745356_Cat_7_Patient Care Record (Inpatien...	866972291_Cat_6_Patient Care Record (Inpatien...	867167556_Cat_4_Patient Care Record (Inpatien...
867167556_Cat_4_Patient Care Record (Inpatien...	867773642_Cat_7_Patient Care Record (Inpatien...	867773642_Cat_7_Patient Care Record (Inpatien...	867808171_Cat_4_Patient Care Record (Inpatien...	867808171_Cat_4_Patient Care Record (Inpatien...
868194296_Cat_5_Patient Care Record (Inpatien...	868194296_Cat_5_Patient Care Record (Inpatien...	868528977_Cat_5_Patient Care Record (Inpatien...	868528977_Cat_5_Patient Care Record (Inpatien...	868533852_Cat_5_Patient Care Record (Inpatien...
868533852_Cat_5_Patient Care Record (Inpatien...	868589632_Cat_ 6-7_Patient Care Record (Inpat...	868589632_Cat_ 6-7_Patient Care Record (Inpat...	868981160_Cat_6_Patient Care Record (Inpatien...	868981160_Cat_6_Patient Care Record (Inpatien...
869060647_Cat_4_Patient Care Record (Inpatien...	869060647_Cat_4_Patient Care Record (Inpatien...	869095984_Cat_6_Patient Care Record (Inpatien...	869240003_Cat_7_Patient Care Record (Inpatien...	869240003_Cat_7_Patient Care Record (Inpatien...
869643579_Cat_7_Patient Care Record (Inpatien...	869643579_Cat_7_Patient Care Record (Inpatien...	869797158_Cat_8_Patient Care Record (Inpatien...	869797158_Cat_8_Patient Care Record (Inpatien...	872566043_Cat_5_Patient Care Record (Inpatien...
872566043_Cat_5_Patient Care Record (Inpatien...	872871395_Cat_5_Patient Care Record (Inpatien...	872871395_Cat_5_Patient Care Record (Inpatien...	872877037_Cat_6_Patient Care Record (Inpatien...	872877037_Cat_6_Patient Care Record (Inpatien...
875352977_Cat_ 4-5_Patient Care Record (Inpat...	875352977_Cat_ 4-5_Patient Care Record (Inpat...	879996159_Cat_ 5-6_Patient Care Record (Inpat...	879996159_Cat_ 5-6_Patient Care Record (Inpat...	882668111_Cat_8_Patient Care Record (Inpatien...
882668111_Cat_8_Patient Care Record (Inpatien...	882695046_Cat_7_Patient Care Record (Inpatien...	882754748_Cat_4_Patient Care Record (Inpatien...	882754748_Cat_4_Patient Care Record (Inpatien...	883054456_Cat_7_Patient Care Record (Inpatien...
883054456_Cat_7_Patient Care Record (Inpatien...	883716700_Cat_6_Patient Care Record (Inpatien...	883918817_Cat_4_Patient Care Record (Inpatien...	883918817_Cat_4_Patient Care Record (Inpatien...	889029982_Cat_6_Patient Care Record (Inpatien...
889029982_Cat_6_Patient Care Record (Inpatien...	892486974_Cat_7_Patient Care Record (Inpatien...			

Handling 2018 New Sample Data from TTSH

ED (Emergency) Notes
Sample Data

Total: **45** files

560277507_Cat_7_ED Notes	561005678_Cat_6_ED Notes	568831565_Cat_6_ED Notes	571192553_Cat_2_ED Notes	575911924_Cat_6_ED Notes
576867505_Cat_6_ED Notes	578935973_Cat_7_ED Notes	579345112_Cat_8_ED Notes	579514165_Cat_5_ED Notes	584121291_Cat_6_ED Notes
585496133_Cat_5_ED Notes	663198575_Cat_7_ED Notes	667598022_Cat_7_ED Notes	668051254_Cat_5_ED Notes	677153773_Cat_6_ED Notes
679020725_Cat_7_ED Notes	862033106_Cat_3-4_ED Notes	862073014_Cat_6_ED Notes	862096578_Cat_4_ED Notes	862096586_Cat_4_ED Notes
862166495_Cat_5_ED Notes	862232907_Cat_6_ED Notes	862254714_Cat_4_ED Notes	862270808_Cat_7_ED Notes	862298725_Cat_2_ED Notes
862323333_Cat_7_ED Notes	862328287_Cat_6_ED Notes	862342662_Cat_5_ED Notes	862345159_Cat_8_ED Notes	862348514_Cat_6_ED Notes
862349405_Cat_6_ED Notes	862358364_Cat_3_ED Notes	862367768_Cat_6_ED Notes	862407657_Cat_6_ED Notes	862408876_Cat_5_ED Notes
862456463_Cat_5_ED Notes	862462779_Cat_5_ED Notes	862465085_Cat_8_ED Notes	862469994_Cat_6_ED Notes	862572175_Cat_5_ED Notes
862649075_Cat_5_ED Notes	862810017_Cat_4_ED Notes	862861782_Cat_5_ED Notes	862875722_Cat_3-4_ED Notes	862895545_Cat_3_ED Notes

IFA Sample Data

Total: **92** files ??

Handling 2018 New Sample Data from TTSH

Patient Care Record Sample Data (consists of Admission & Discharge)

Total: **92** files

- 560277507_Cat_7_Patient Care Record (Inpatien...
- 560277507_Cat_7_Patient Care Record (Inpatien...
- 561005678_Cat_6_Patient Care Record (Inpatien...
- 561005678_Cat_6_Patient Care Record (Inpatien...
- 568831565_Cat_6_Patient Care Record (Inpatien...
- 568831565_Cat_6_Patient Care Record (Inpatien...
- 571192553_Cat_2_Patient Care Record (Inpatien...
- 571192553_Cat_2_Patient Care Record (Inpatien...
- 575911924_Cat_6_Patient Care Record (Inpatien...
- 575911924_Cat_6_Patient Care Record (Inpatien...
- 576867505_Cat_6_Patient Care Record (Inpatien...
- 576867505_Cat_6_Patient Care Record (Inpatien...
- 578935973_Cat_7_Patient Care Record (Inpatien...
- 578935973_Cat_7_Patient Care Record (Inpatien...
- 579345112_Cat_8_Patient Care Record (Inpatien...
- 579345112_Cat_8_Patient Care Record (Inpatien...
- 579514165_Cat_5_Patient Care Record (Inpatien...
- 579514165_Cat_5_Patient Care Record (Inpatien...
- 584121291_Cat_6_Patient Care Record (Inpatien...
- 584121291_Cat_6_Patient Care Record (Inpatien...
- 585496133_Cat_5_Patient Care Record (Inpatien...
- 585496133_Cat_5_Patient Care Record (Inpatien...
- 663198575_Cat_7_Patient Care Record (Inpatien...
- 663198575_Cat_7_Patient Care Record (Inpatien...
- 667598022_Cat_7_Patient Care Record (Inpatien...
- 667598022_Cat_7_Patient Care Record (Inpatien...
- 668051254_Cat_5_Patient Care Record (Inpatien...
- 668051254_Cat_5_Patient Care Record (Inpatien...
- 677153773_Cat_6_Patient Care Record (Inpatien...
- 677153773_Cat_6_Patient Care Record (Inpatien...
- 679020725_Cat_7_Patient Care Record (Inpatien...
- 679020725_Cat_7_Patient Care Record (Inpatien...
- 862033106_Cat_3-4_Patient Care Record (Inpati...
- 862033106_Cat_3-4_Patient Care Record (Inpati...
- 862073014_Cat_6_Patient Care Record (Inpatien...
- 862073014_Cat_6_Patient Care Record (Inpatien...
- 862096578_Cat_4_Patient Care Record (Inpatien...
- 862096578_Cat_4_Patient Care Record (Inpatien...
- 862096586_Cat_4_Patient Care Record (Inpatien...
- 862096586_Cat_4_Patient Care Record (Inpatien...
- 862166495_Cat_5_Patient Care Record (Inpatien...
- 862166495_Cat_5_Patient Care Record (Inpatien...
- 862232907_Cat_6_Patient Care Record (Inpatien...
- 862232907_Cat_6_Patient Care Record (Inpatien...
- 862254714_Cat_4_Patient Care Record (Inpatien...
- 862254714_Cat_4_Patient Care Record (Inpatien...
- 862270808_Cat_7_Patient Care Record (Inpatien...
- 862270808_Cat_7_Patient Care Record (Inpatien...
- 862298725_Cat_2_Patient Care Record (Inpatien...
- 862298725_Cat_2_Patient Care Record (Inpatien...
- 862323333_Cat_7_Patient Care Record (Inpatien...
- 862323333_Cat_7_Patient Care Record (Inpatien...
- 862328287_Cat_6_Patient Care Record (Inpatien...
- 862328287_Cat_6_Patient Care Record (Inpatien...
- 862342662_Cat_5_Patient Care Record (Inpatien...
- 862342662_Cat_5_Patient Care Record (Inpatien...
- 862345159_Cat_8_Patient Care Record (Inpatien...
- 862345159_Cat_8_Patient Care Record (Inpatien...
- 862348514_Cat_6_Patient Care Record (Inpatien...
- 862348514_Cat_6_Patient Care Record (Inpatien...
- 862349405_Cat_6_Patient Care Record (Inpatien...
- 862349405_Cat_6_Patient Care Record (Inpatien...
- 862358364_Cat_3_Patient Care Record (Inpatien...
- 862358364_Cat_3_Patient Care Record (Inpatien...
- 862367768_Cat_6_Patient Care Record (Inpatien...
- 862367768_Cat_6_Patient Care Record (Inpatien...
- 862407657_Cat_6_Patient Care Record (Inpatien...
- 862407657_Cat_6_Patient Care Record (Inpatien...
- 862408876_Cat_5_Patient Care Record (Inpatien...
- 862408876_Cat_5_Patient Care Record (Inpatien...
- 862456463_Cat_5_Patient Care Record (Inpatien...
- 862456463_Cat_5_Patient Care Record (Inpatien...
- 862462779_Cat_5_Patient Care Record (Inpatien...
- 862462779_Cat_5_Patient Care Record (Inpatien...
- 862465085_Cat_8_Patient Care Record (Inpatien...
- 862465085_Cat_8_Patient Care Record (Inpatien...
- 862469994_Cat_6_Patient Care Record (Inpatien...
- 862469994_Cat_6_Patient Care Record (Inpatien...
- 862572175_Cat_5_Patient Care Record (Inpatien...
- 862572175_Cat_5_Patient Care Record (Inpatien...
- 862649075_Cat_5_Patient Care Record (Inpatien...
- 862649075_Cat_5_Patient Care Record (Inpatien...
- 862810017_Cat_4_Patient Care Record (Inpatien...
- 862810017_Cat_4_Patient Care Record (Inpatien...
- 862861782_Cat_5_Patient Care Record (Inpatien...
- 862861782_Cat_5_Patient Care Record (Inpatien...
- 862875722_Cat_3-4_Patient Care Record (Inpati...
- 862875722_Cat_3-4_Patient Care Record (Inpati...
- 862894859_Cat_6_Patient Care Record (Inpatien...
- 862894859_Cat_6_Patient Care Record (Inpatien...
- 862895545_Cat_3_Patient Care Record (Inpatien...
- 862895545_Cat_3_Patient Care Record (Inpatien...

Difference between Old and New Sample Data

Old Sample Data

- Documents have been converted before we handle them
- Word documents' data format is inconsistent
- Example of inconsistency in data format:
  - Page breaks
  - Missing values in tables

New Sample Data

- Documents were in PDF and not converted
- Word documents have to be converted to PDF files
- Newer sample data format is more consistent

Timeline (Week 1-5) - Lynn

Week 1

- Learning of Python & MongoDB
- Retrieving ED Notes from MongoDB
- Understanding text mining

Week 2

- Writing a program to extract informative key words (Unigram, Bigram & Trigram)
- Removing stop words using stopwords list

Week 3

- Retrieving of ED Notes to store in an array for text analysis later
- Improving Unigram, Bigram & Trigram

Week 4

- Retrieving ED Notes in a format of "cat_id | content"
- Amending codes in docs2txt_source.py to convert all ED Notes MS Words into text files
- Creating a method in ednotes_extractor.py that returns an array (array to store ednotes)

Week 5

- Trying to finish up docs2txt_source.py for ED Notes
- Combine my two testing projects together into one project

Timeline (Week 6-11) - Lynn

Week 6-7

- Create a program that handles 3 different input files (ED Notes, Patient Care Records Admission, Patient Care Records Discharge)
- Create inpatientcare extractor to extract Patient Care Records Admission and Discharge

Week 8

- N-grams (Knowledge-based) - Using this for keyword matching
- Handling new and old data sets for oversampling to improve the mean accuracy of data sets

Week 9

- Trying another approach to improve the mean accuracy of data sets
- Learning of Topic Modeling and trying the practical for topic modeling

Week 10

- Trying to implement topic modeling into the current working project
- Create keywords_consensus_analysis.py to do keyword matching

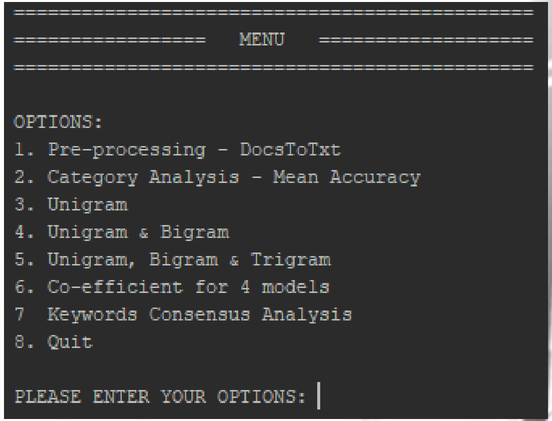
Week 11

- Extract word terms from LDA (Latent Dirichlet Allocation) model in topic modeling
- Store extracted word terms in a dictionary and append to ngram key phrases list

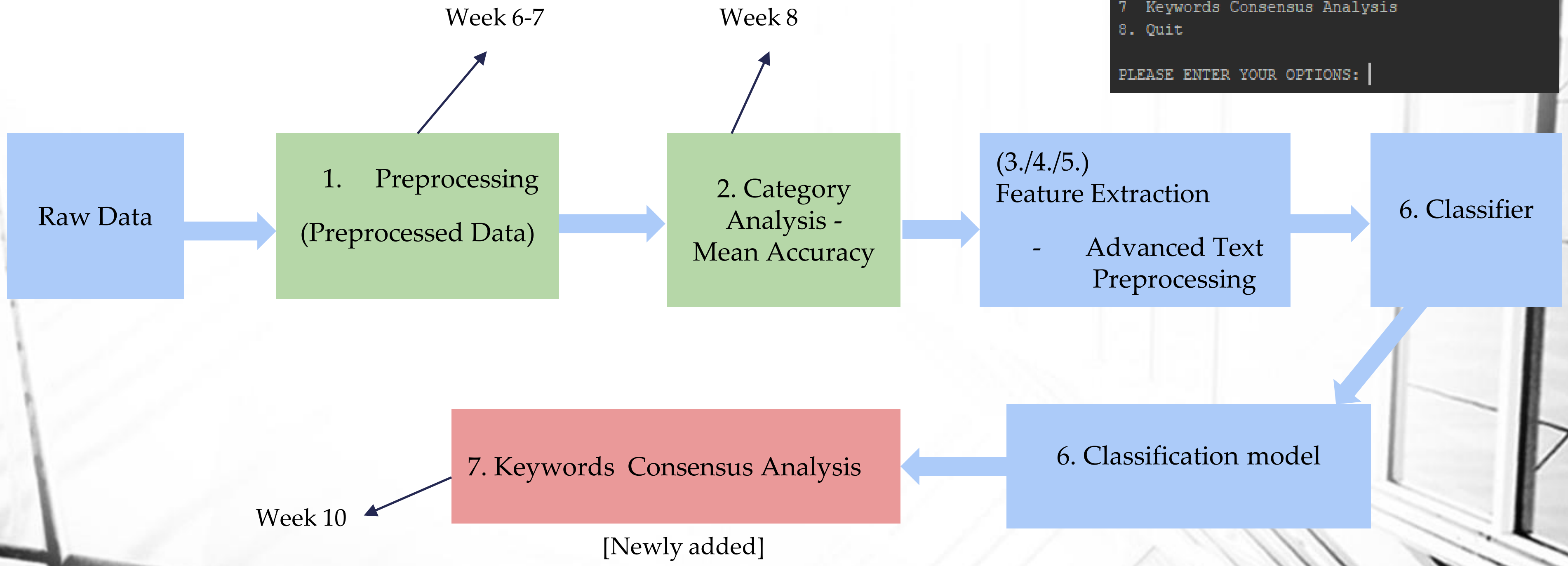
Tasks Accomplished - Lynn

- Create inpatientcare_extractor.py to extract Patient Care Record Discharge and Admission data directly from word documents.
- Create a program to handle 3 different input files (ED Notes, Patient Care Record_Admission & Patient Care Record_Discharge) for text analysis.
- Improve mean accuracy results by using oversampled data (Oversampling of imbalanced data)
- Create key_consensus_analysis.py to do keyword matching
- Create topicmodel_keywords_extractor.py to extract keywords from LDA (Latent Dirichlet Allocation) model in topic modeling

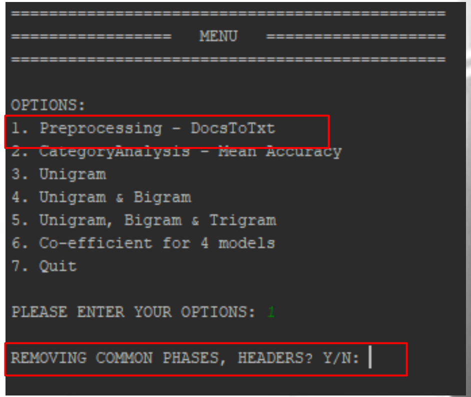
Flowchart of Project - Lynn



```
==============================================
================    MENU    ==================
==============================================

OPTIONS:
1. Pre-processing - DocsToTxt
2. Category Analysis - Mean Accuracy
3. Unigram
4. Unigram & Bigram
5. Unigram, Bigram & Trigram
6. Co-efficient for 4 models
7  Keywords Consensus Analysis
8. Quit

PLEASE ENTER YOUR OPTIONS: |
```



Flowchart of Project - Lynn



ED Notes, Patient Care Record Admission, Patient Care Record Discharge

Raw Data

1. Preprocessing

(Preprocessed Data)

Week 6-7

Basic preprocessing of text data

Go to docs2txt_source.py that handles all 3 different output files (ED Notes, Patient Care Record Admission & Patient Care Record Discharge)

Yes

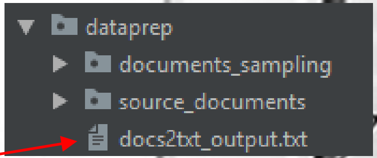
Remove common phases/headers

No

Method - get_ednotes(source_directory) in ednotes_extractor.py called by docs2txt_source.py

Goes to source directory, gets each ed notes file, stores in an array & it returns an array

Save all output files into docs2txt_output.txt file in “cat_id|content” format



Flowchart of Project - Lynn

```
==================================================
=================     MENU     ===================
==================================================

OPTIONS:
1. Preprocessing - DocsToTxt
2. CategoryAnalysis - Mean Accuracy
3. Unigram
4. Unigram & Bigram
5. Unigram, Bigram & Trigram
6. Co-efficient for 4 models
7. Quit

PLEASE ENTER YOUR OPTIONS: 2
```

1. Preprocessing

(Preprocessed Data)

2. Category Analysis - Mean Accuracy

Week 8

Category Analysis - Mean Accuracy

The method - mean_acc() in the category_analysis.py is being called

```
model_name
LinearSVC                 0.340540
LogisticRegression        0.351332
MultinomialNB             0.330268
RandomForestClassifier    0.298286
Name: accuracy, dtype: float64
```

***Note: Oversampling for old Patient Care Records Admission and Discharge only

Before Oversampling:

model_name	
LinearSVC	0.350736
LogisticRegression	0.360346
MultinomialNB	0.351688
RandomForestClassifier	0.399394

After Oversampling:

model_name	
LinearSVC	0.757010
LogisticRegression	0.751377
MultinomialNB	0.615189
RandomForestClassifier	0.690137

Flowchart of Project - Lynn

```
================================================
================   MENU   ================
================================================

OPTIONS:
1. Preprocessing - DocsToTxt
2. CategoryAnalysis - Mean Accuracy
3. Unigram
4. Unigram & Bigram
5. Unigram, Bigram & Trigram
6. Co-efficient for 4 models
7. Quit

PLEASE ENTER YOUR OPTIONS: 5
```

2. Category Analysis - Mean Accuracy

3./4./5. Feature Extraction

- Advanced Text Preprocessing

```
'''ngram_range(1, 1) -> unigram / ngram_range(1, 3) -> unigram, bigram, trigram'''
tfidf = TfidfVectorizer(sublinear_tf=True, min_df=1, norm='l2', encoding='UTF-8', ngram_range=(1, 3),
                        stop_words=stop_words)
```

N-grams:

- Unigram - extract a single word
- Bigram - extract a word pair
- Trigram - extract triple words

For example: ngram_range(1, 1) means unigram, ngram_range(1, 3) means unigram, bigram and trigram

Flowchart of Project - Lynn

```
=============================================
=================   MENU   ==================
=============================================

OPTIONS:
1. Preprocessing - DocsToTxt
2. CategoryAnalysis - Mean Accuracy
3. Unigram
4. Unigram & Bigram
5. Unigram, Bigram & Trigram
6. Co-efficient for 4 models
7. Quit

PLEASE ENTER YOUR OPTIONS: 6
```

3./4./5. Feature Extraction

- Advanced Text Preprocessing

6. Classifier

6. Classification model

4 Classifiers:

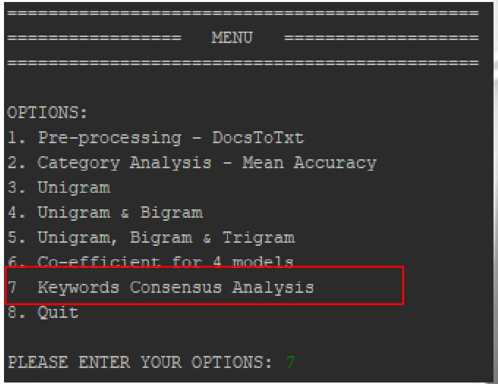
- Naive Bayes (Multinomial)
- Logistic Regression
- Linear SVC
- Random Forest

```
RUNNING CATEGORY ANALYSIS...

SELECT A MODEL TYPE:
1. Naive Bayes
2. Logistic Regression
3. Linear SVC
4. Random Forest

PLEASE ENTER YOUR OPTION:
```

Classifier	Classification Model
An algorithm that maps input data to a specific category	Draws some conclusion from input values given for training

Flowchart of Project - Lynn

Objective: Predict category from an unknown document string



7. Keywords Consensus Analysis

Method - check_keyword( ) in keywords_consensus_analysis.py

1. Read an unknown document in a text file as a string

2. Read key_phases text file, which consists of ngram keywords as a dictionary

3. Compare unknown document string with the ngram keywords dictionary

4. If keywords comparison found and match, then add count

The total count is then divided by the total number of keywords in the dictionary.

5. Return the highest float value and the predicted category

Flowchart of Project - Lynn

```
===============================================
=================    MENU    ==================
===============================================

OPTIONS:
1. Pre-processing - DocsToTxt
2. Category Analysis - Mean Accuracy
3. Unigram
4. Unigram & Bigram
5. Unigram, Bigram & Trigram
6. Co-efficient for 4 models
7  Keywords Consensus Analysis
8. Quit

PLEASE ENTER YOUR OPTIONS: 7
```

Objective: Predict category from an unknown document string

7. Keywords Consensus Analysis

```python
def check_keyword():
```

```python
# read unknown text from text file as a string
with open(unknown, "r") as fp:
    unknown_file = fp.read()
```

```python
# read key phases from text file as a dictionary
df = pd.read_csv(key_phrases, sep='|')
key_phases_dict = df.to_dict(orient='records')
```

```python
# compare content of unknown file with key phases (write a for loop to do keyword matching)
for key in new_dict.keys():
    count_dict[key] = 0
    new_list2 = new_dict[key].split(",")
    new_dict[key] = new_list2
    for j in new_dict[key]:
        if j in unknown_file:
            # if word found and matched, add count
            count_dict[key] = count_dict[key] + 1
    count_dict[key] = float(count_dict[key] / len(new_list2))
```

```python
count = max(count_dict.values())
# get the maximum value in the dictionary
max_value = [(k, v) for k, v in count_dict.items() if v == count]
maxval = print(max_value)
```

7. Keywords Consensus Analysis

Objective: Predict category from an unknown document string

```
# compare content of unknown file with key phases (write a for loop to do keyword matching)
for key in new_dict.keys():
    count_dict[key] = 0
    new_list2 = new_dict[key].split(",")
    new_dict[key] = new_list2
    for j in new_dict[key]:
        if j in unknown_file:
            # if word found and matched, add count
            count_dict[key] = count_dict[key] + 1
    count_dict[key] = float(count_dict[key] / len(new_list2))
print(count_dict)
```

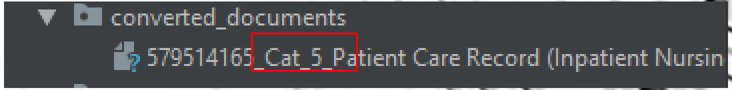
```
{2: 0.02666666666666667, 3: 0.006666666666666667, 4: 0.013333333333333334, 5: 0.04, 6: 0.08, 7: 0.03333333333333333, 8: 0.013333333333333334}
```

```
count = max(count_dict.values())
# get the maximum value in the dictionary
max_value = [(k, v) for k, v in count_dict.items() if v == count]
maxval = print(max_value)
```

```
[(6, 0.08)]
```

Predicted document to be category 6

Unknown document string is category 5



Flowchart of Project - Lynn

Objective: To add more informative words into key_phrases.txt to improve the prediction of an unknown document string

7. Keywords Consensus Analysis

```
==================================================
=================    MENU    =================
==================================================

OPTIONS:
1. Pre-processing - DocsToTxt
2. Category Analysis - Mean Accuracy
3. Unigram
4. Unigram & Bigram
5. Unigram, Bigram & Trigram
6. Co-efficient for 4 models
7  Keywords Consensus Analysis
8. Quit

PLEASE ENTER YOUR OPTIONS: 7
```

Method - get_topicmodel_words (filename) in topicmodel_keywords_extractor.py

Read in each category text file as a filename

Extract keywords from each category from LDA Model in topic modeling

key_phrases to match cat_id with the extracted word terms

If match then append to the words to the string of the particular category

```
get_topicmodel_words(output_cat3)
get_topicmodel_words(output_cat4)
get_topicmodel_words(output_cat5)
get_topicmodel_words(output_cat6)
get_topicmodel_words(output_cat7)
get_topicmodel_words(output_cat8)
```

Problems Encountered - Lynn

- To handle new data sets given to us
- To handle old data sets with inconsistency in data format
- To improve on the Ngram's keyword extraction results

Solutions - Lynn

- Seek help from previous batch students and supervisor
- Research online for solutions (e.g. stackoverflow)
- Trial and error by using several different methods
    - Different methods meaning to try out and test if online related codes to the project works

Summing-up - Lynn

docs2txt_source.py

(for handling 3 different output files - ED Notes, Patient Care Record Admission & Patient Care Record Discharge)

To save all output files into docs2txt_output.txt file in "cat_id|content" format

ednotes_extractor.py

(for extracting ED Notes directly from the word documents)

To extract all ED Notes data text from the word documents

inpatientcare_extractor

(for extracting Patient Care Records Admission & Discharge directly from the word documents)

Extract method using docx2txt.process( )

import docx2txt: to extract text from docx files

keywords_consensus_analysis.py

(for doing keyword matching between an unknown document string and a known dictionary of words)

To predict category from unknown document string

topicmodel_keywords_extractor.py

(for extracting keywords for each category from LDA Model in Topic Modeling)

To add on keywords to the key_phrases text file, which consists of ngram key phrases for each category

Timeline (Week 1-5) - Ren Ern

Week 1

- Exploring and learning about Python language

Week 2

- Made use of Acrobat Action Wizard to convert PDF files to MS Word

Week 3

- Improving on previous student's IFA extract to extract data from MS Word Documents

Week 4

- Improving on previous student's IFA extract to extract data from MS Word Documents

Week 5

- Creating a method in the extract to read the array and convert it into a text file
- Tried to store variables in text file in a binomial format

Timeline (Week 6-11) - Ren Ern

Week 6-7

- Converting Yes/No from IFA files to binomial format

Week 8

- Converting of IFA Word Document files to raw text format

Week 9

- Performing Text Analysis on Raw Output Text

Week 10

- Converting Yes/No from IFA files to binomial format
- Building a Classifier and Predictor

Week 11

- Improving on the accuracy on the Classifier and Predictor

Tasks Accomplished - Ren Ern

- Converting of IFA documents from PDF to Word Document using Adobe Acrobat
- Improving on the previous batch student's IFA extract to extract more data
- Converting IFA documents from Word Document format to Raw Output Text format
- Converting Yes/No from IFA files and converting them to binomial format
- Create a Classifier to identify relationships and patterns among the categories
- Create a Predictor to predict a category base on the Classifier

Introduction

- IFA files contain medical information about a patient

- They are split into 5 categories

1. Sensory and Communications
2. Functional Status
3. Mental Health
4. Other Geriatric Syndromes
5. Social

DEPARTMENT OF GERIATRIC MEDICINE
Inpatient Frailty Assessment (IFA)

CHIEF COMPLAINT:

SENSORY and COMMUNICATION		
Visual Impairment	Yes	Both
Hearing Impairment	Yes	Both
Communication	Yes	Verbal

FUNCTIONAL STATUS	
Decline in function compared to premorbid	No

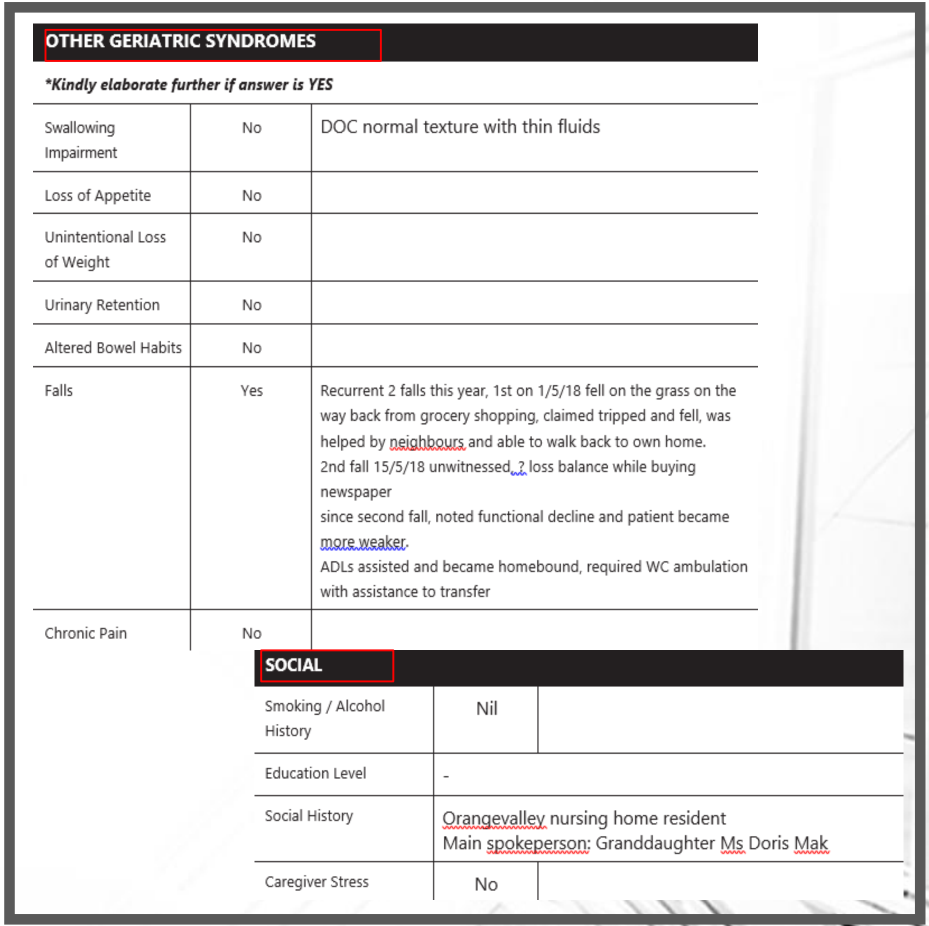
MENTAL HEALTH		
Progressive Forgetfulness or Known Cognitive impairment	Yes	Advanced Dementia
Depression	Nil	
Sleep Issues	Yes	Sleep- wake reversal
Behavioral Disturbance	Nil	
Cognitive - known Advanced Dementia		

Introduction

- IFA files contain medical information about a patient

- They are split into 5 categories

1. Sensory and Communications
2. Functional Status
3. Mental Health
4. Other Geriatric Syndromes
5. Social



OTHER GERIATRIC SYNDROMES

***Kindly elaborate further if answer is YES**

Swallowing Impairment	No	DOC normal texture with thin fluids
Loss of Appetite	No	
Unintentional Loss of Weight	No	
Urinary Retention	No	
Altered Bowel Habits	No	
Falls	Yes	Recurrent 2 falls this year, 1st on 1/5/18 fell on the grass on the way back from grocery shopping, claimed tripped and fell, was helped by neighbours and able to walk back to own home. 2nd fall 15/5/18 unwitnessed, ? loss balance while buying newspaper since second fall, noted functional decline and patient became more weaker. ADLs assisted and became homebound, required WC ambulation with assistance to transfer
Chronic Pain	No	

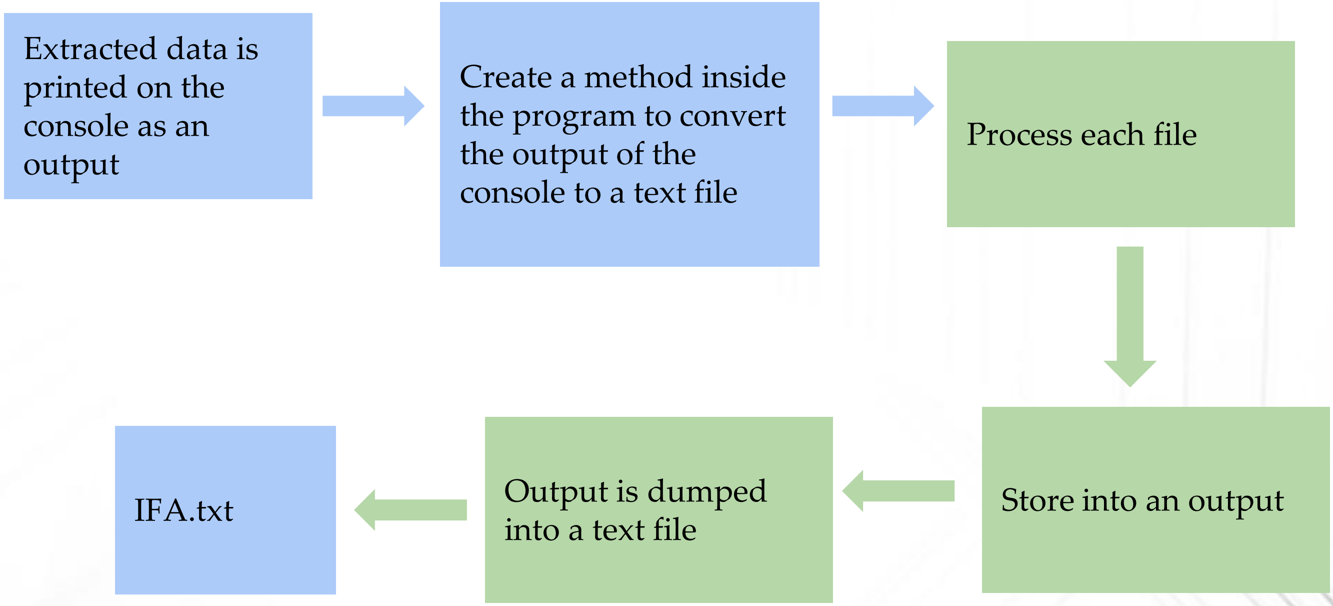
SOCIAL

Smoking / Alcohol History	Nil	
Education Level	-	
Social History	Orangevalley nursing home resident Main spokeperson: Granddaughter Ms Doris Mak	
Caregiver Stress	No	

Flowchart of Project - Ren Ern

1. Use of Acrobat Action Wizard to convert IFA files in PDF to Word Document format
2. IFA word document files is passed through the program
3. Based on certain formats of codes, relevant data will be extracted
4. Added conditions for columns that were missing
5. Creates an exception, so that when faced with an error, the program is able to handle it
6. This creates a dictionary, which is then converted to a JSON object to be returned
7. a JSON object is created with a key and assigns an empty dictionary to it
8. Prints and returns the JSON object created from the IFA file passing through the program
9. Extracted data is printed on the console as an output

Flowchart of Project - Ren Ern



Flowchart of Project - Ren Ern

IFA.txt

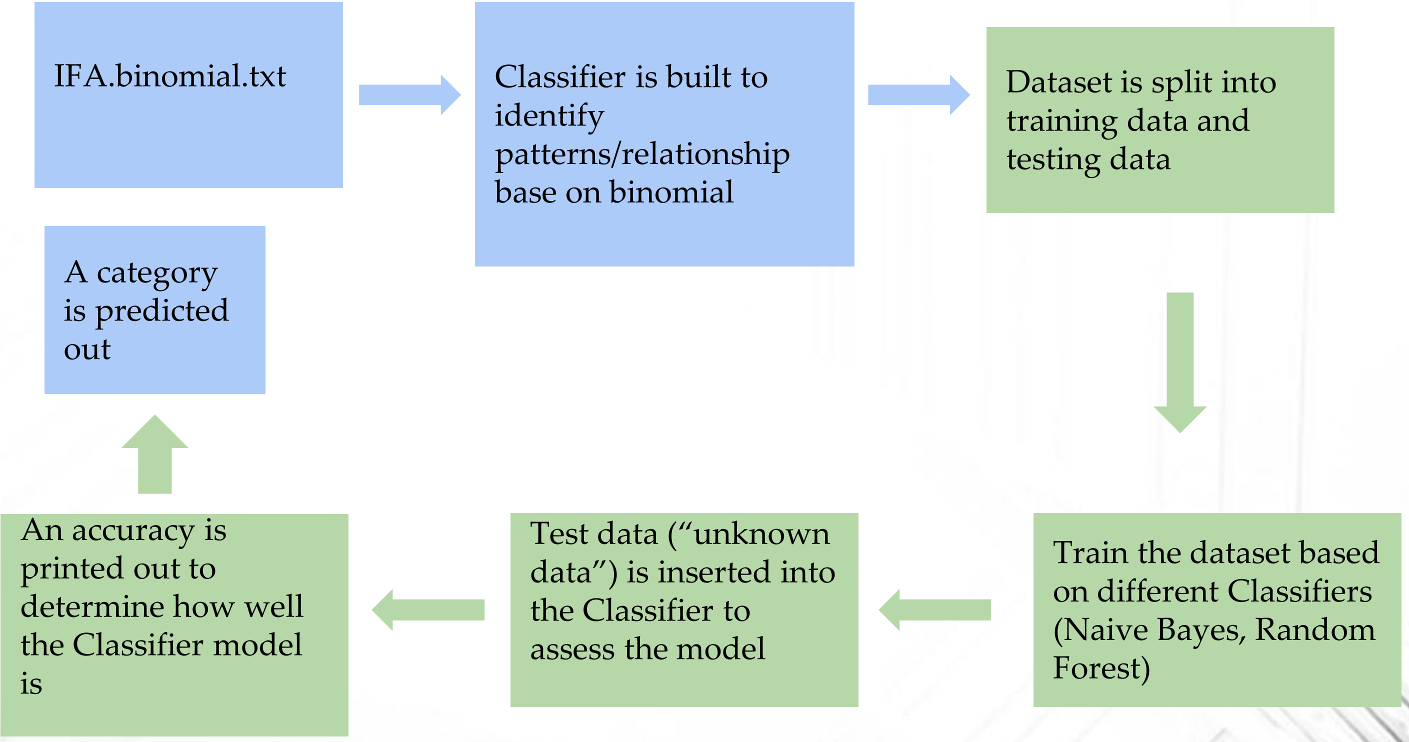
Converting of Yes/No to binomial format

Based on certain keywords, program extracts out Yes/No and converts it to 1/0 respectively

If there are any missing values, program will set it as -1

IFA.binomial.txt

Flowchart of Project - Ren Ern



Problems Encountered - Ren Ern

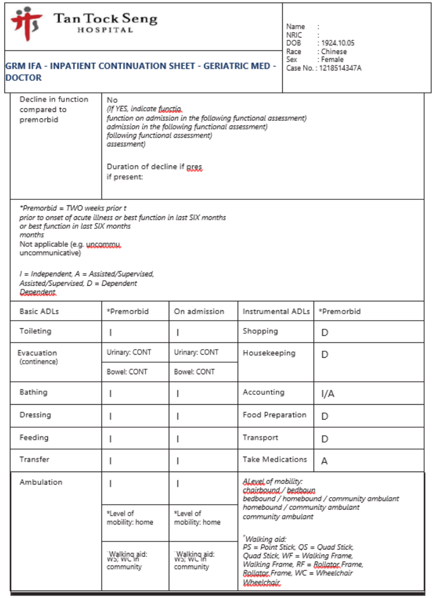
- Previous student’s extract was hard-coded to a single document, it was not “flexible” in reading different format of IFA documents

```
Run: IFA_Extract
C:\Users\L31308\AppData\Local\Programs\Python\Python37\python.exe C:/Users/L31308/Desktop/backup/IFA__Extract.py
C:/Users/L31308/Desktop/FYP/Word Documents/860246760_Cat_5_IFA.docx
C:/Users/L31308/Desktop/FYP/Word Documents/860305644_Cat_5_IFA.docx
C:/Users/L31308/Desktop/FYP/Word Documents/860389965_Cat_5_IFA.docx
C:/Users/L31308/Desktop/FYP/Word Documents/860512095_Cat_5_IFA.docx
Traceback (most recent call last):
  File "C:/Users/L31308/Desktop/backup/IFA__Extract.py", line 923, in <module>
    format(str2)
  File "C:/Users/L31308/Desktop/backup/IFA__Extract.py", line 575, in format
    ,'Toileting, Premorbid':functionallist[2],'Toileting, Admission':functionallist[3],
IndexError: list index out of range
```

Only 4 files were able to be read.

Problems Encountered - Ren Ern

No page breaks, program was able to read


Tan Tock Seng HOSPITAL

GRM IFA - INPATIENT CONTINUATION SHEET - GERIATRIC MED - DOCTOR

Name :
NRIC
DOB : 1924.10.05
Race : Chinese
Sex : Female
Case No. : 1218514347A

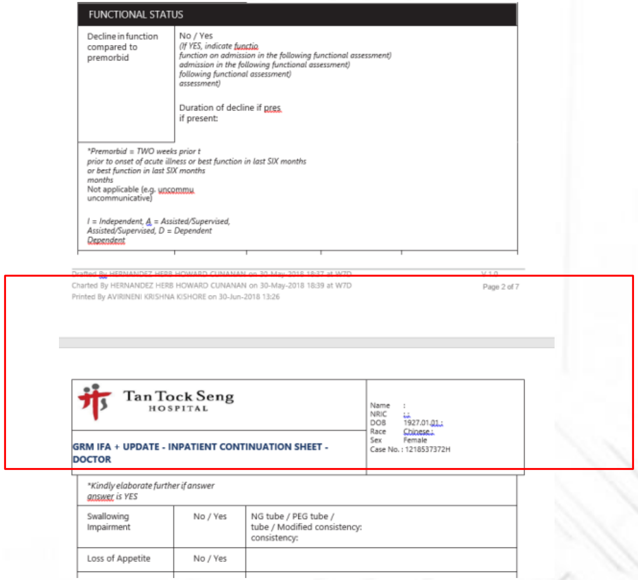
Decline in function compared to premorbid	No (If YES, indicate functio. function on admission in the following functional assessment) admission in the following functional assessment) following functional assessment) assessment) Duration of decline if pres if present:

*Premorbid = TWO weeks prior t
prior to onset of acute illness or best function in last SIX months
or best function in last SIX months
months
Not applicable (e.g. uncommu.
uncommunicative)

I = Independent, A = Assisted/Supervised,
Assisted/Supervised, D = Dependent
Dependent

Basic ADLs	*Premorbid	On admission	Instrumental ADLs	*Premorbid
Toileting	I	I	Shopping	D
Evacuation (continence)	Urinary: CONT Bowel: CONT	Urinary: CONT Bowel: CONT	Housekeeping	D
Bathing	I	I	Accounting	I/A
Dressing	I	I	Food Preparation	D
Feeding	I	I	Transport	D
Transfer	I	I	Take Medications	A
Ambulation	I *Level of mobility: home Walking aid: WS; WC in community	I *Level of mobility: home Walking aid: WS; WC in community	ALevel of mobility: chairbound / bedboun bedbound / homebound / community ambulant homebound / community ambulant community ambulant ^Walking aid: PS = Point Stick, QS = Quad Stick, Quad Stick, WF = Walking Frame, Walking Frame, RF = Rollator Frame, Rollator Frame, WC = Wheelchair Wheelchair	

Previous student's extract was unable to read this document due to the page break


FUNCTIONAL STATUS

Decline in function compared to premorbid	No / Yes (If YES, indicate functio function on admission in the following functional assessment) admission in the following functional assessment) following functional assessment) assessment) Duration of decline if pres. if present:

*Premorbid = TWO weeks prior t
prior to onset of acute illness or best function in last SIX months
or best function in last SIX months
months
Not applicable (e.g. uncommu.
uncommunicative)

I = Independent, A = Assisted/Supervised,
Assisted/Supervised, D = Dependent
Dependent

Charted By HERNANDEZ HERB HOWARD CUNANAN on 30-May-2018 18:39 at W7D Page 2 of 7
Printed By AVIRINENI KRISHNA KISHORE on 30-Jun-2018 13:26

Tan Tock Seng HOSPITAL

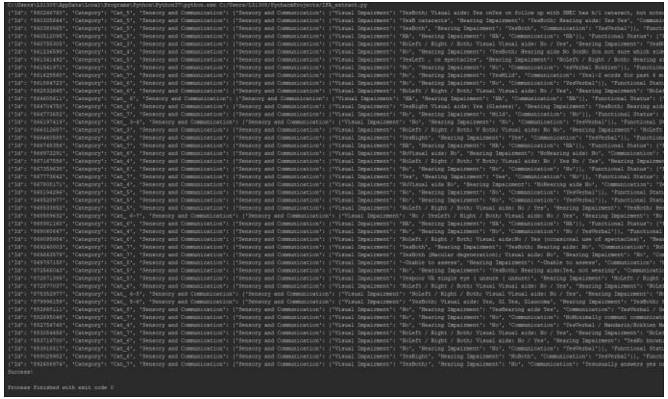
GRM IFA + UPDATE - INPATIENT CONTINUATION SHEET - DOCTOR

Name :
NRIC :
DOB 1927.01.01
Race Chinese
Sex Female
Case No. : 1218537372H

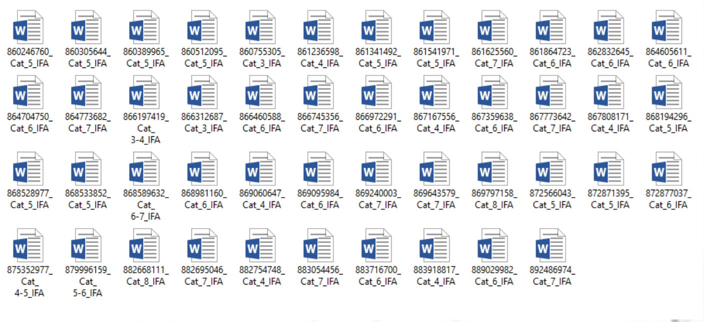
*Kindly elaborate further if answer
answer is YES

Swallowing Impairment	No / Yes	NG tube / PEG tube / tube / Modified consistency: consistency:
Loss of Appetite	No / Yes	

Solutions - Ren Ern

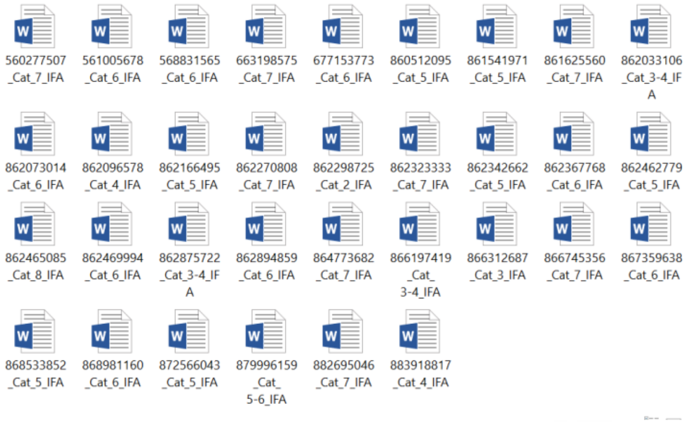


Contents of Word Document printed on the console



Program was able to read 60 out of 90 documents

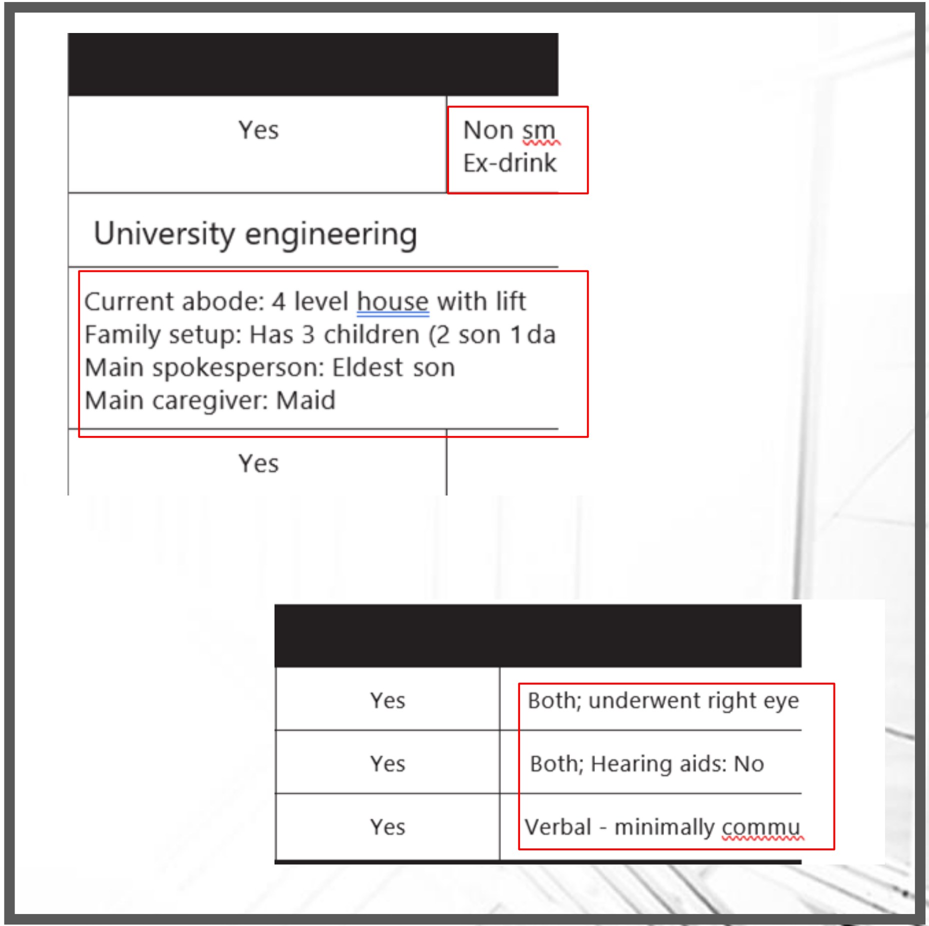
Problems Encountered- Ren Ern



Documents that were still unable to be read

Problems Encountered- Ren Ern

- Half of the Word Documents that were unable to be extracted had tables that were "cut"
- Hence program was not able to read the data
- 16 out of 30 documents had "broken"/"cut" tables


[redacted]

Yes	Non sm Ex-drink
University engineering	
Current abode: 4 level house with lift Family setup: Has 3 children (2 son 1 da Main spokesperson: Eldest son Main caregiver: Maid	
Yes	

[redacted]

Yes	Both; underwent right eye
Yes	Both; Hearing aids: No
Yes	Verbal - minimally commu

Problems Encountered- Ren Ern

- Program extracts data by hitting certain keywords already pre-set in the codes
- Some files had empty list
- I went to open up the files to take a look, and I found out that they had different names
- Hence program was unable to catch that part of the data
- Added in that particular naming (ie, SENSORY and COMMUNI COMMUNICATION) for program to read and extract data

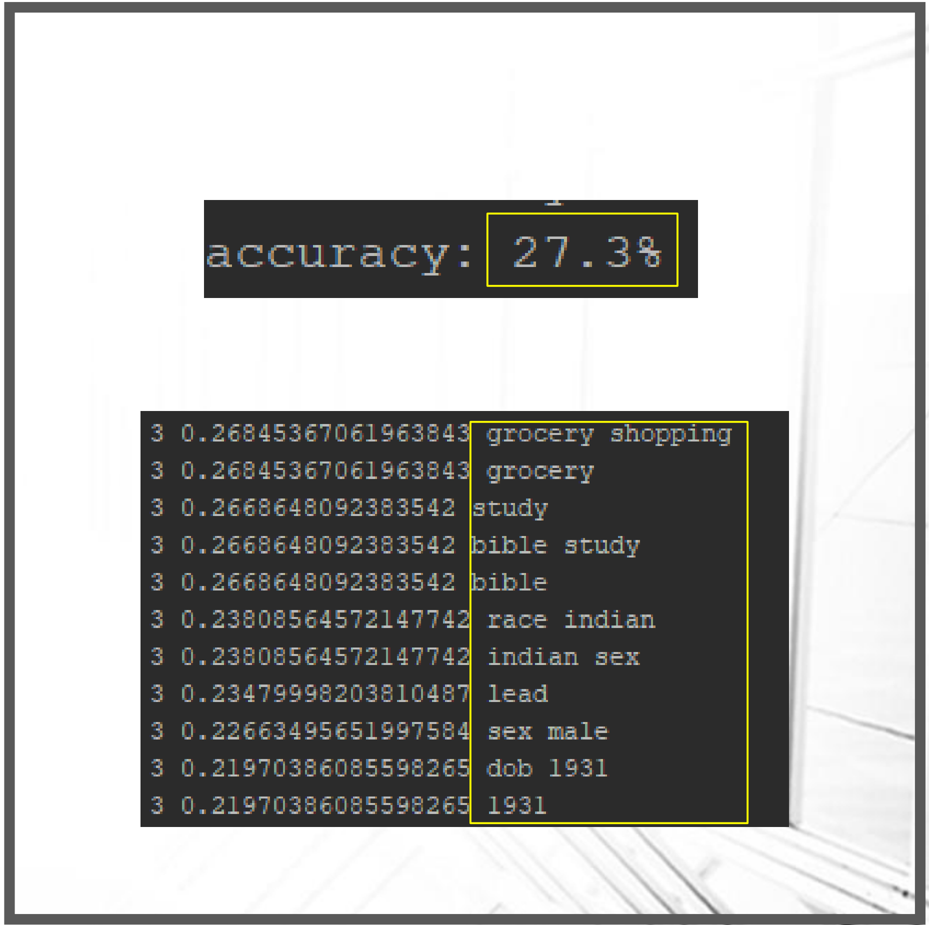
```
{'Sensory and Communication': {'Vis
': {'Sensory and Communication': {'V
{'Sensory and Communication': {'Vis
': {'Sensory and Communication': {'V
{'Sensory and Communication': {'Vis
{'Sensory and Communication': {'Vis
{'Sensory and Communication': {}},
```

SENSORY and COMMUNI COMMUNICATION	
Visual Impairment	no
Hearing Impairment	yes
Communication	Yes

```
for paral in cell.paragraphs:
    if (paral.text == 'SENSORY and COMMUNI COMMUNICATION' or paral.text == 'SENSORY and COMMUNICATION'
        sector = 'SENSORY and COMMUNICATION'
    else:
        pass
```

Problems Encountered- Ren Ern

- Tried to perform keyword extraction, but accuracy was too low
- Keywords were irrelevant too



Problems Encountered- Ren Ern

MENTAL HEALTH		
Progressive Forgetfulness or Known Cognitive impairment	No	noted Hx mild cognitive cognitive impairment s/b Dr Selva in 2014 noted to have STML (+) with occasional repetition and apraxia (+), (+) with occasional repetition and apraxia (+), but still able to repetition and apraxia (+), but still able to pack angpao apraxia (+), but still able to pack angpao able to pack angpao AMT 6/10 (no schooling) schooling) discharged from GRM GRM
Depression	No	Suicide risk: No looks f looks forward to weekends as grandchildren visit weekends as grandchildren visit grandchildren visit
Sleep Issues	Yes	wakes up at 8am every mor every morning reported she can't sleep well, wakes up x2 nightly. no nocturia sleep well, wakes up x2 nightly. no nocturia (wears diapers at x2 nightly. no nocturia (wears diapers at night due to falls risk) nocturia (wears diapers at night due to falls risk) diapers at night due to falls risk) to falls risk)
Behavioral Disturbance	No	personality: very particula particular, likes things to be neat and tidy things to be neat and tidy and tidy

File with comments

MENTAL HEALTH		
Progressive Forgetfulness or Known Cognitive impairment	NA	
Depression	No	
Sleep Issues	No	
Behavioral Disturbance	No	

File with no comments

IFA files are not suitable for text analysis as not every file has texts made by the doctor

Problems Encountered- Ren Ern

SENSORY and COMMUNI COMMUNICATION		
Visual Impairment	no	0
Hearing Impairment	yes	1
Communication	Yes	1

```
cat_3, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
cat_4, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
cat_4-5, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,
cat_6-7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
cat_2, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
cat_3, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0,
```

- IFA.binomial program reads Yes/No and convert it to 1/0 respectively. If it's neither Yes or No, for example, null values/missing values the program sets it to -1
- Some files were very obvious outliers, the whole dataset was set to -1

Problems Encountered- Ren Ern

MENTAL HEALTH	
Progressive Forgetfulness or Known Cognitive impairment	Y
Depression	N
Sleep Issues	Y
Behavioral Disturbance	N

```
if (theline[0:2]=="no" or theline[0:1] == "n"):
    output += "0"
elif (theline[0:3]=="yes" or theline[0:1] == "y"):
    output += "1"
else:
    output += "-1"
```

- I went to investigate and opened up the files to take a look and some files indicated Yes/No by only stating either Y/N
- Added in some codes for program to accommodate and still convert to binomial format

Problems Encountered- Ren Ern

SENSORY and COMMUNICATION	
Visual Impairment	**No** / Yes
Hearing Impairment	**No** / Yes
Communication	No / **Yes**

SENSORY and COMMUNICATION	
Visual Impairment	**No**
Hearing Impairment	**No**
Communication	**Yes**

- I went to investigate and opened up the files to take a look and some files indicated Yes/No by bolding the character, hence program set it as -1
- I changed the dataset myself, and only kept the bolded character

Problems Encountered- Ren Ern

```
C:\Users\renern\Anaconda3\python.exe C:/Users/renern/Desktop/FYP/IFA_classifier.py
the accuracy of the Naive bayes classifier on the test set is : 40.0 %
the accuracy of the Random forest classifier on the test set is : 26.666666666666668 %

Process finished with exit code 0
```

- Performance of the classifier depends heavily on the data quality
- Accuracy was too low as dataset was too little

Problems Encountered- Ren Ern

```
the accuracy of the Naive bayes classifier on the test set is : 58.97435897435898 %
the accuracy of the Random forest classifier on the test set is : 64.1025641025641 %

Process finished with exit code 0
```

- Did oversampling to try to improve accuracy
- Accuracy was able to increase

Problems Encountered- Ren Ern

- Predictor predicted out Category but it wasn't accurate since the accuracy of the classifier wasn't that great to begin with
- Performance of Classifier depends heavily on data quality
- We had too little data and dataset was repetitive due to oversampling
- It would not result in a fair and accurate result

```
The real category is : 4
The predicted category with RandomForest is : [5]
The predicted category with NaiveBayes is : [7]
```

```
The real category is : 6
The predicted category with RandomForest is : [5]
The predicted category with NaiveBayes is : [4]
```

```
The real category is : 7
The predicted category with RandomForest is : [5]
The predicted category with NaiveBayes is : [4]
```

Summing-up - Ren Ern

IFA_json_extract.py
(for extracting data out from IFA files)

A method was created in the extract to run it in the Classifier when called

→

IFA_binomial.py
(for converting Yes/Nos to 1/0s)

A method was created in the script to run it in the Classifier when called

→

IFA_classifier.py
(to build a model and predict categories)

Classifier runs for both **IFA_json_extract.py** and **IFA_binomial**.py when program is run


Q&A

Thank You for Listening!