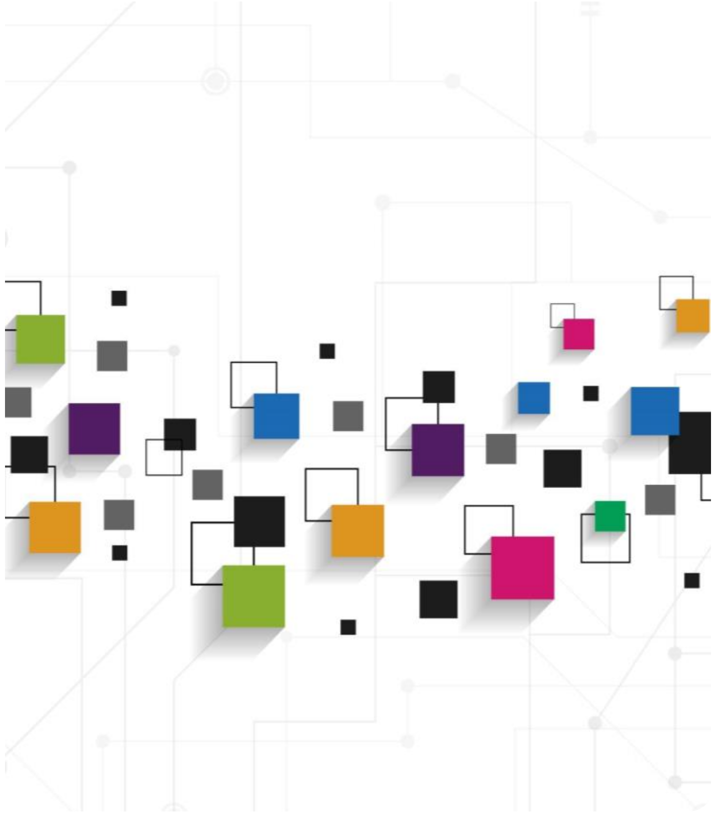
Preprocessing

Ramaseshan Ramachandran

High level overview

- Introduction to NLP
- Counting and empirical laws
- Word Representation using a vector
- Finding Word vectors –
  Artificial Neural Network approach
- Language Generation
- Language Models Using Deep ANN
- Language Translation
- Transformers
- Chat-bots, Question Answering Systems



Is NLP Hard?

- Multiple ways of representing the same scenario
- Includes common sense and contextual representation
- Complex representation information (simple to hard vocabulary)
- Mixing of visual cues
- Ambiguous in nature
- Idioms, metaphors, sarcasm (Yeah! right), double negatives, etc. make it difficult for automatic processing
- Human language interpretation depends on real world, common sense, and contextual knowledge

Corpus Creation

Gather possible problem statements

Collect problems related to every possible domain , if possible

Convert them into textual (ascii) form

Copy them into a folder

Sample Corpus Content

An airplane accelerates down a runway at 3.20 m/s2 for 32.8 s until is finally lifts off the ground. Determine the distance traveled before takeoff.

Ben Rushin is waiting at a stoplight. When it finally turns green, Ben accelerated from rest at a rate of a 6.00 m/s2 for a time of 4.10 seconds. Determine the displacement of Ben's car during this time period.

A train brakes from 40 m/s to a stop over a distance of 100 m. a) What is the acceleration of the train? b) How much time does it take the train to stop?

The space shuttle releases a space telescope into orbit around the earth. The telescope goes from being stationary to traveling at a speed of 1700 m/s in 25 seconds. What is the acceleration of the satellite?

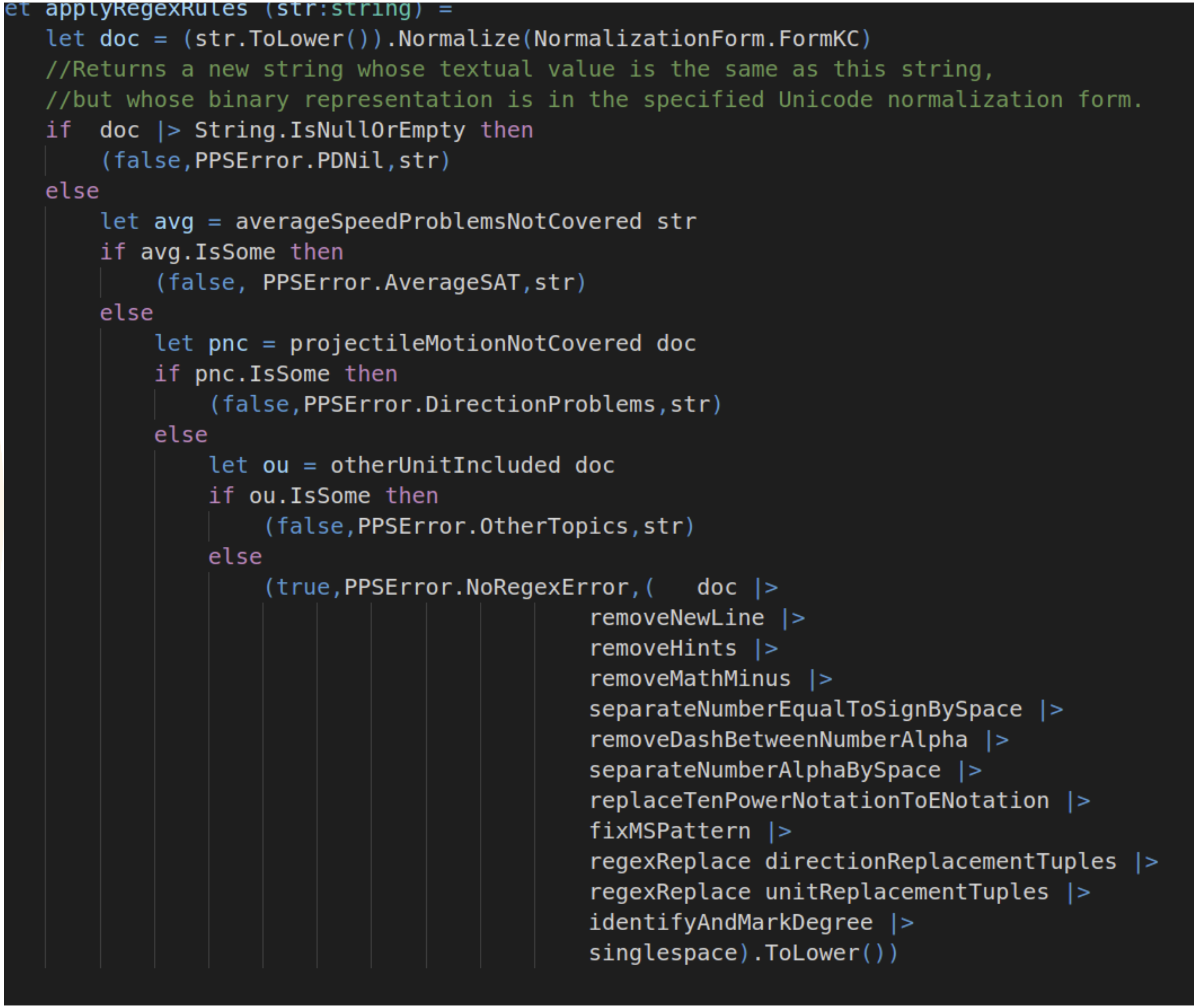
Preprocessing

- Make the corpus accessible through standard APIs
    - CSV, PDF, WORD, XML, JSON, XLS, etc.
- A good understanding of the corpus is required
- Any modification required for processing?
    - HTML2Txt, PDF2Txt,WORD2Txt
    - Removing unwanted(?) words from text
    - Convert document to ascii text format (may lose information)
    - Case folding (Converting all text to small letters)
    - Remove foreign characters and words

Improper preprocessing schemes may lead to loss of lexical context

Preprocessing steps are unique to a problem

Preprocessing - Specific To Word problems



```
et applyRegexRules (str:string) =
    let doc = (str.ToLower()).Normalize(NormalizationForm.FormKC)
    //Returns a new string whose textual value is the same as this string,
    //but whose binary representation is in the specified Unicode normalization form.
    if  doc |> String.IsNullOrEmpty then
        (false,PPSError.PDNil,str)
    else
        let avg = averageSpeedProblemsNotCovered str
        if avg.IsSome then
            (false, PPSError.AverageSAT,str)
        else
            let pnc = projectileMotionNotCovered doc
            if pnc.IsSome then
                (false,PPSError.DirectionProblems,str)
            else
                let ou = otherUnitIncluded doc
                if ou.IsSome then
                    (false,PPSError.OtherTopics,str)
                else
                    (true,PPSError.NoRegexError,(   doc |>
                                            removeNewLine |>
                                            removeHints |>
                                            removeMathMinus |>
                                            separateNumberEqualToSignBySpace |>
                                            removeDashBetweenNumberAlpha |>
                                            separateNumberAlphaBySpace |>
                                            replaceTenPowerNotationToENotation |>
                                            fixMSPattern |>
                                            regexReplace directionReplacementTuples |>
                                            regexReplace unitReplacementTuples |>
                                            identifyAndMarkDegree |>
                                            singlespace).ToLower())
```

COMMONLY USED PREPROCESSING STEPS FOR ENGLISH

Preprocessing consists of (a) tokenization, (b) normalization and (c) substitution

Case folding - Convert all text into lower case

Stemming - running → run

Lemmatization - best → good

Remove misspellings

Punctuations, white space, newline, tabs...

- Removing contractions -
  - isn't → is not, I'd → i would
- Remove scripts, form variables, - HTML and XML
- Tokenization
- ….
- ….

Preprocessing ...

Original Corpus

An airplane accelerates down a runway at 3.20m/s2 for 32.8 s until it finally liftsoff the ground. Determine the distance traveled before takeoff

How far will a car travel in 25 min at 12 kmh?

A jalopy with an initial speed of 23.7 km/h accelerates at a uniform rate of 0.92m/s2Â for 3.6 s. Find the final speed and the displacement of the jalopy during this time

Processed Corpus

- An airplane accelerates down a runway at 3.20 < m/s * *2 > for 32.8 < s > until it finally lifts off the ground. Determine the distance traveled before takeoff
- How far will a car travel in 25 < minute > at 12 < km/h >?
- A jalopy with an initial speed of 23.7 < km/h > accelerates at a uniform rate of 0.92 < m/s * *2 > for 3.6 < s >. Find the final speed and the displacement of the jalopy during this time

HTML Preprocessing

- Convert HTML to Text – Remove all marks ups <HTML>, </HTML>, <BODY> <P> <FORM>
- Remove scripts
- Normalize text
- …
- …

Problems with Preprocessing

- Case folding removes contextual information
- Regex converts unwanted information
- Standard tokenization are white space based
- ….
- ….
- ….
- ….

Human/Machine Learning

- How do we solve problems when we lack sufficient knowledge?
- Finding Examples and using experience gained are useful
- Examples provide certain underlying patterns
- Patterns give the ability to predict some outcome or help in constructing an approximate model
- The model may help resolve some problems, though may not be an ideal one
- If learning is the key, what and how do we do we learn?
- What learning algorithm do we use?
- Can we use it to train the machine?

Regular Expression

Definition

Regular Expression defines a pattern representing a class of strings in a document

Examples

- A regular expression representing a year pattern would fetch a set of years from a document, if available
  - A simple example would be \d{4} or [0-9]{4}
- How does one capture 0000-2999 and not 3000+?
  - How does one capture dates 1-31 and not 32+?
  - How does one capture month, 1-12, and not 13+?
  - How does one capture dd-mm-yyyy patterns?
- How does one match Aadhaar number in a running text?

Regex Grammar

? mark matches zero or one of the previous expression

. Matches any character

[0-9]+ matches one or more digits

[a-z]* matches any sequence of strings

R[a-z] matches a string that starts with R

\sR[a-z]* - What does this regex match?

\w matches [0-9a-zA-Z]* all words\

\W matches all [^0-9a-zA-Z]

Write a regex to find **all patterns that end with *ing***

Harder Examples

- Find all representations of numbers
    - Integers, float, +10e+12, -34.45, 5x10**2, etc.
- Replace ten_power_notation With e_notation in a document
- Remove hints
    - Example Hint - (g=9.8m/s)
- Match citations from a research paper
    - Example [12][13] [12,13] [Ada 2021] [Ada and Pad 2021]
- Match emails in documents

Test your understanding

Select the sorted list of a set of words extracted from the documents (given below) by this regular expression

(^[AD][a-z]+)|(\s[aD]+[a-z]+)

- Around the world in 8 dollars
- Bags are not free any more
- Why is Dan doing this?
- Dogs are good at picking the scent of Dora

Pick the right option

a) Around, any, are, Dan, Dogs, dollars, world

b) any, at, Bags, Dan, Dogs, dollars, Dora

c) any, are, Bags, doing, dollars, Dora, good

d) any, are, are, Around, at, Dan, Dogs, Dora

e) any, are, Around, at, Dan, Dogs, Dora

f) None of the above