COMP5046
Natural Language Processing

Lecture 5: Assignment1 and Language Fundamental

Dr. Caren Han

Semester 1, 2022
School of Computer Science,
University of Sydney

0 LECTURE PLAN

Lecture 5: Assignment1 and Language Fundamental

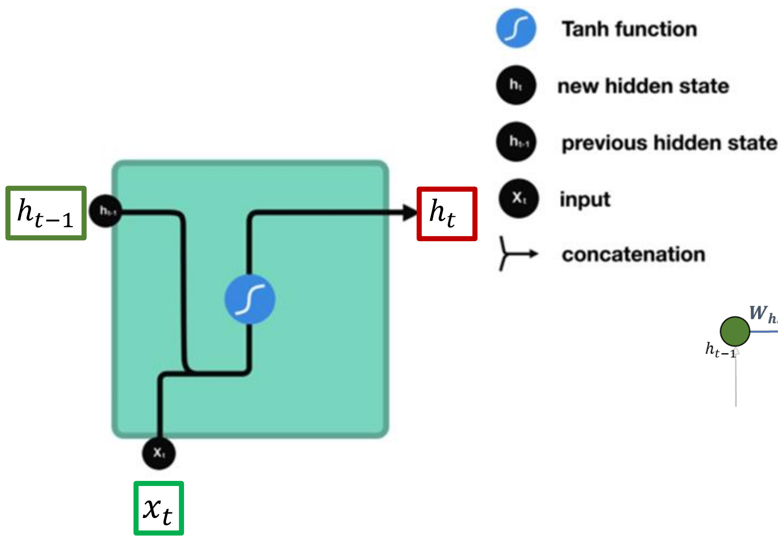
1. **RNN/LSTM, Dealing Context Review**
2. Assignment 1 Discussion
3. Sentiment Analysis
   1. Sentiment Analysis
   2. Sentiment Analysis: Examples
   3. Sentiment Analysis: Lexicons
4. Language Fundamental
   - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
   1. Tokenization
   2. Cleaning and Normalisation
   3. Stemming and Lemmatisation
   4. Stopword
   5. Regular Expression

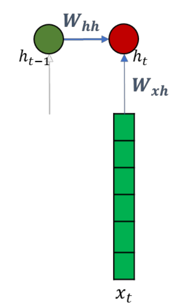
1 RNN/LSTM Review

Neural Network + Memory = Recurrent Neural Network

Hidden Layer

Input Layer



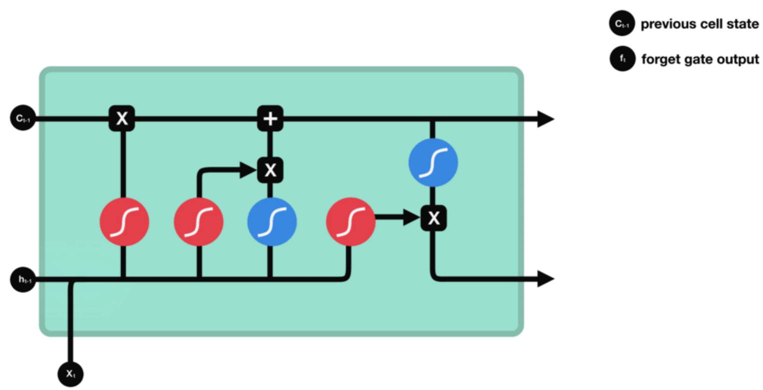


$$\boxed{h_t} = \boxed{tanh}(\underline{W_{hh}}\,\underline{h_{t-1}} + \underline{W_{xh}}\,\underline{x_t} + \underline{b_h})$$

New hidden state *A function with parameters W* *Previous state* *input*

1 RNN/LSTM Review

LSTM (Long Short-Term Memory) – Forget Gate

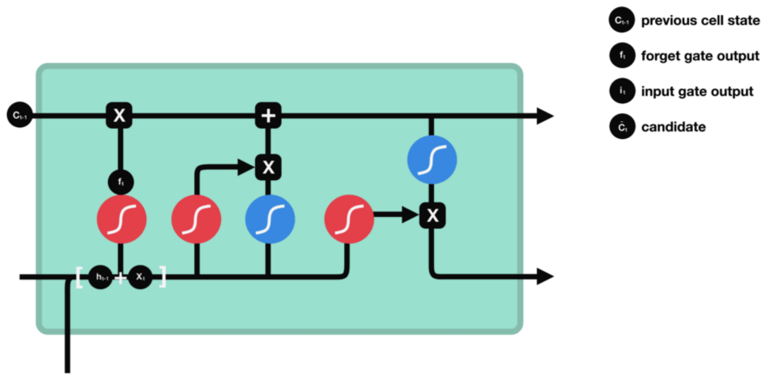


$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

Decides what information should be thrown away or kept

Information from the **previous hidden state** and information from the **current input** is passed through the **sigmoid function**. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep.

LSTM (Long Short-Term Memory) – Input Gate



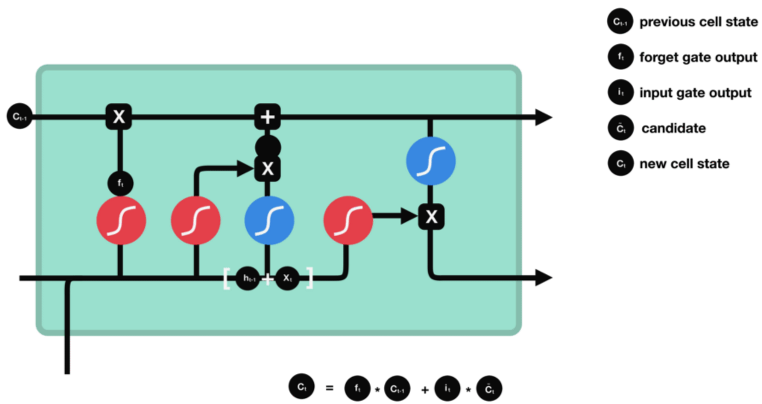
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$$

1. *Pass the previous hidden state and current input into a sigmoid function*
2. *Pass the hidden state and current input into the tanh function to squish values between -1 and 1 to help regulate the network*
3. *Multiply the tanh output with the sigmoid output*

**sigmoid output will decide which information is important to keep from the tanh output*

LSTM (Long Short-Term Memory) – Cell States

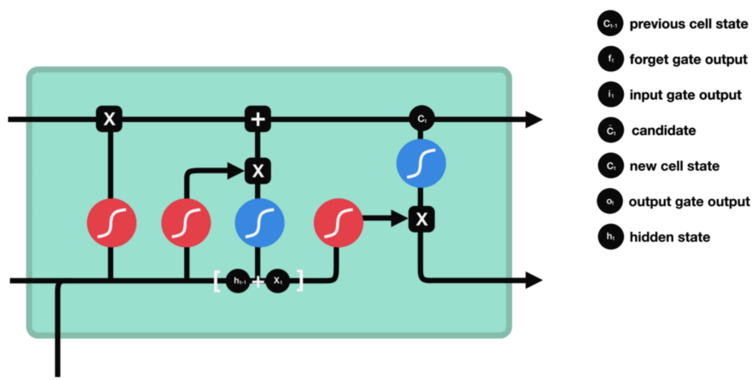


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- *the cell state gets pointwise multiplied by the forget vector*
- *take the output from the input gate and do a pointwise addition which updates the cell state to new values that the neural network finds relevant*
- *That gives us our new cell state*

1 RNN/LSTM Review

LSTM (Long Short-Term Memory) – Output Gate



$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

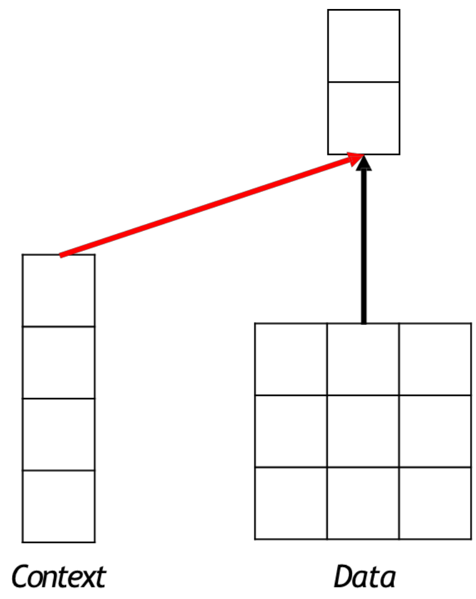
$$h_t = o_t * \tanh(C_t)$$

decides what the next hidden state should be.

- pass the previous hidden state and the current input into a sigmoid function
- pass the newly modified cell state to the tanh function
- multiply the tanh output with the sigmoid output to decide what information the hidden state should carry

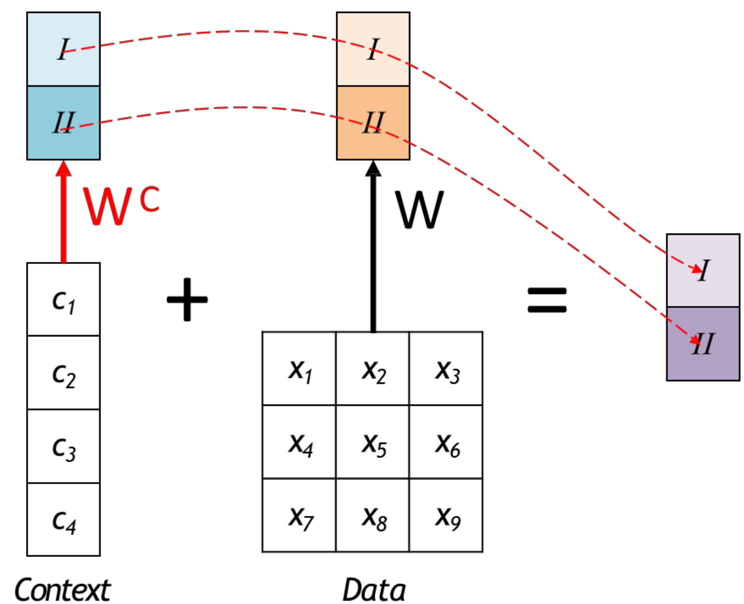
1 Dealing Context: Review

V to V' – Projection with Context (1)



1 Dealing Context: Review

V to V' – Projection with Context (2)



1 Dealing Context: Review

V to V' with Context - Linear Algebra

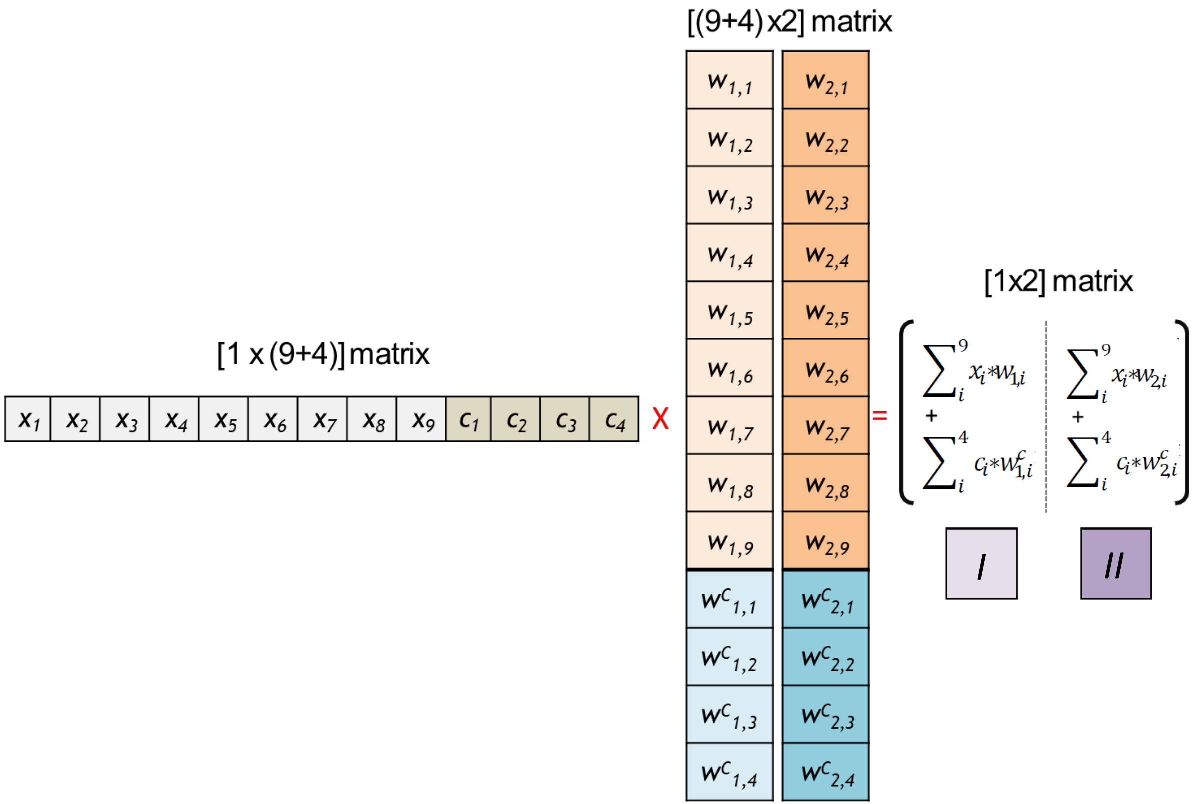
[1 x 9] matrix, [9x2] matrix, [1x2] matrix

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{bmatrix} \times \begin{bmatrix} w_{1,1} & w_{2,1} \\ w_{1,2} & w_{2,2} \\ w_{1,3} & w_{2,3} \\ w_{1,4} & w_{2,4} \\ w_{1,5} & w_{2,5} \\ w_{1,6} & w_{2,6} \\ w_{1,7} & w_{2,7} \\ w_{1,8} & w_{2,8} \\ w_{1,9} & w_{2,9} \end{bmatrix} = \begin{bmatrix} \underbrace{\sum_i^9 x_i * w_{1,i}}_{I} , & \underbrace{\sum_i^9 x_i * w_{2,i}}_{II} \end{bmatrix}$$

[1 x 4] matrix, [1x2] matrix

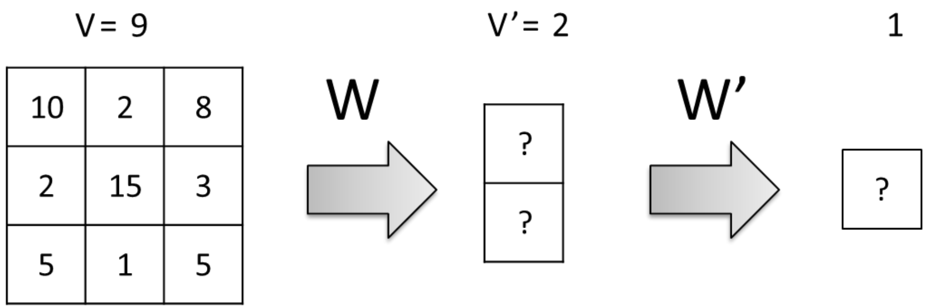
$$\begin{bmatrix} c_1 & c_2 & c_3 & c_4 \end{bmatrix} \times \begin{bmatrix} w^C_{1,1} & w^C_{2,1} \\ w^C_{1,2} & w^C_{2,2} \\ w^C_{1,3} & w^C_{2,3} \\ w^C_{1,4} & w^C_{2,4} \end{bmatrix} = \begin{bmatrix} \underbrace{\sum_i^4 c_i * w^c_{1,i}}_{I} , & \underbrace{\sum_i^4 c_i * w^c_{2,i}}_{II} \end{bmatrix}$$

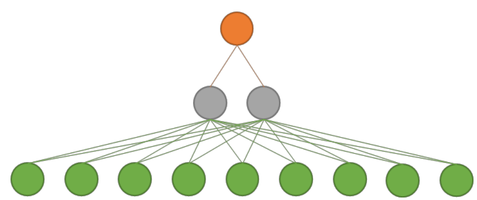
V to V’ with Context - Linear Algebra (Simplified)



1 Dealing Context: Review

V → V' → 1





0

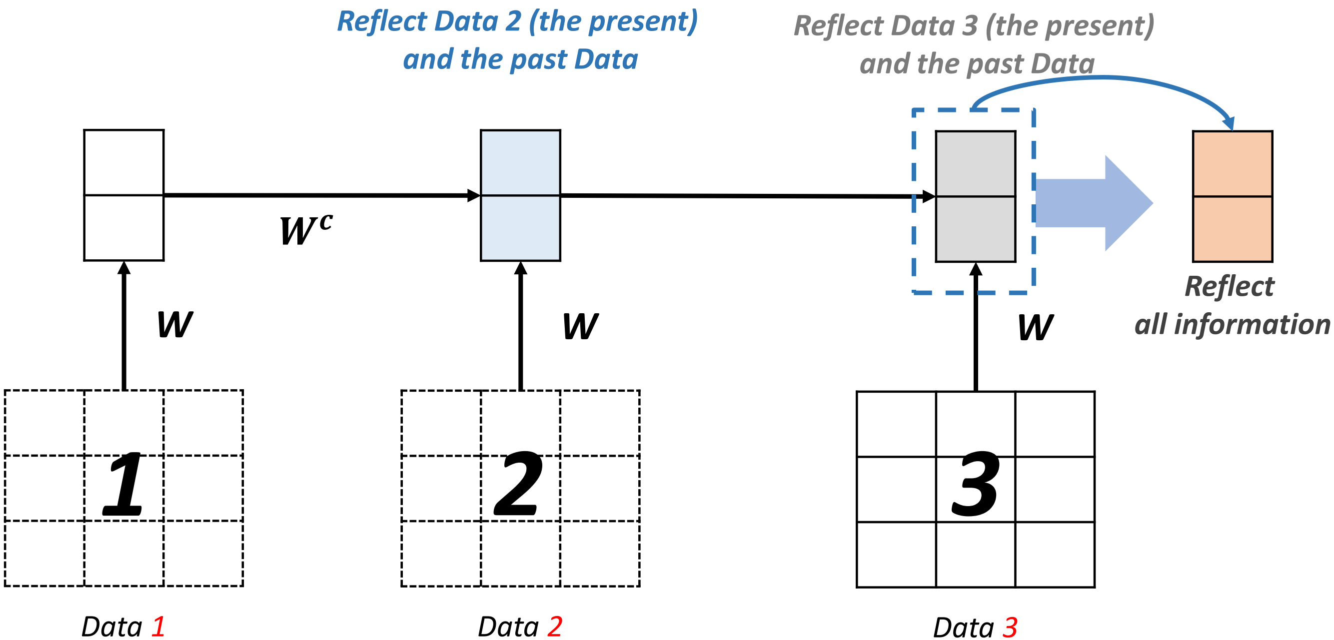
LECTURE PLAN

Lecture 5: Assignment1 and Language Fundamental

1. RNN/LSTM, Dealing Context Review
2. **Assignment 1 Discussion**
3. Sentiment Analysis
    1. Sentiment Analysis
    2. Sentiment Analysis: Examples
    3. Sentiment Analysis: Lexicons
4. Language Fundamental
    - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
    1. Tokenization
    2. Cleaning and Normalisation
    3. Stemming and Lemmatisation
    4. Stopword
    5. Regular Expression

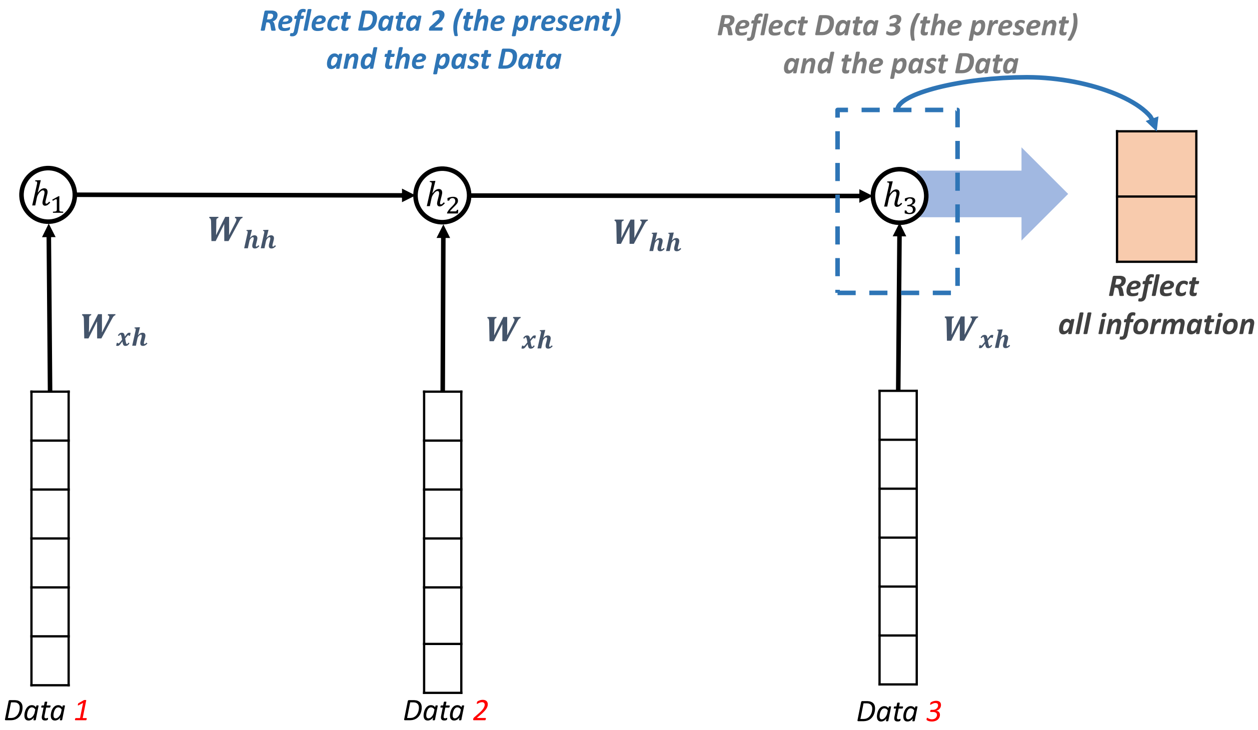
Assignment 1 Discussion

Vs → V's → V'



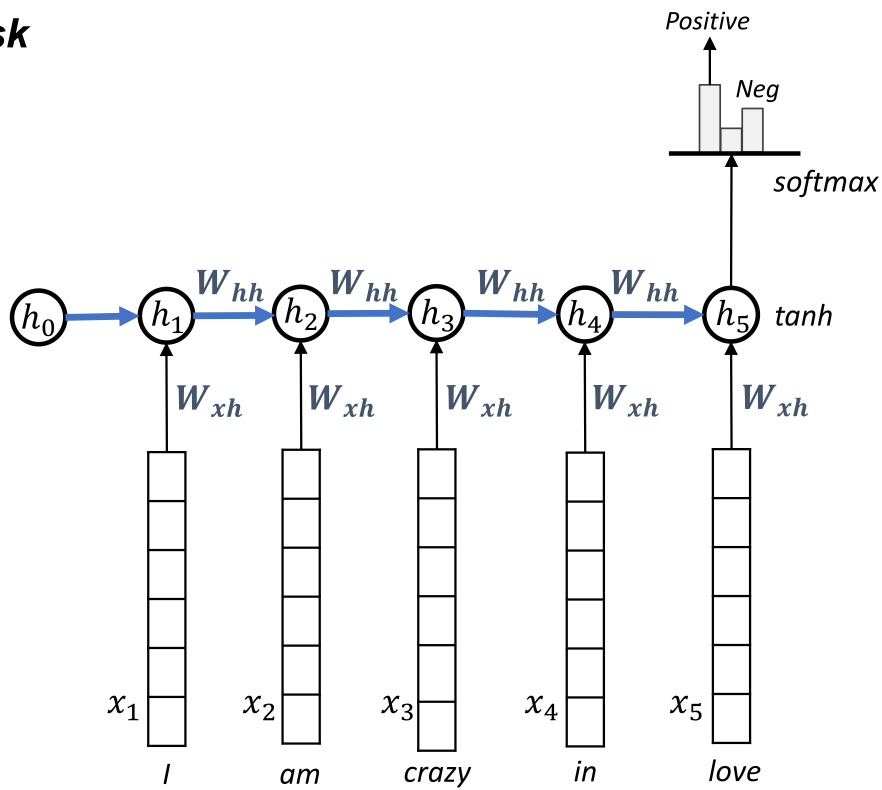
2 Assignment 1 Discussion

Vs → V's → V'



RNN

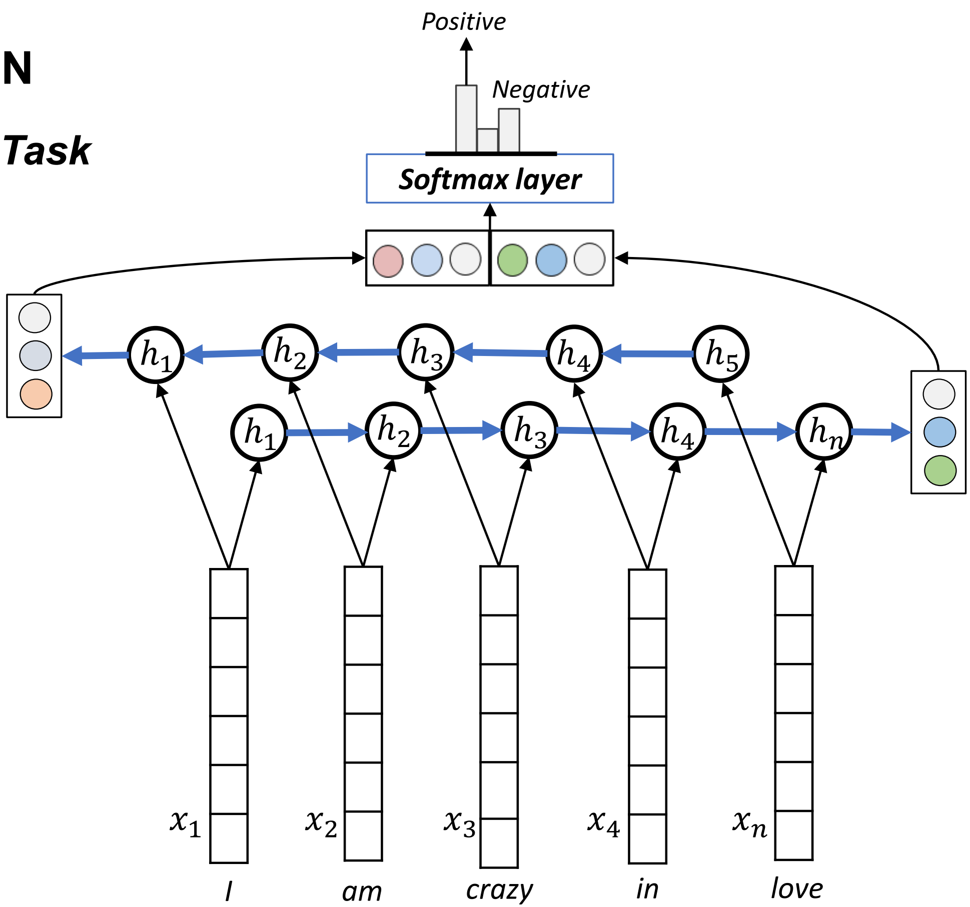
N to 1 Task



2 Assignment 1 Discussion

Bi-RNN

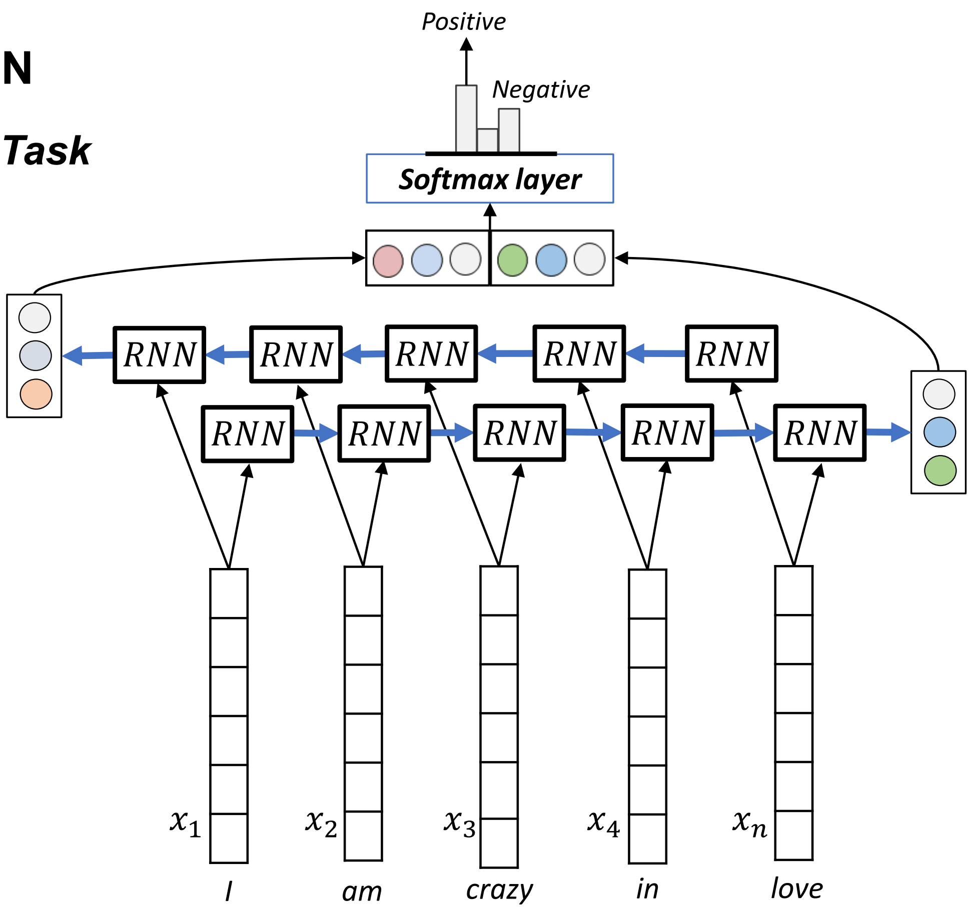
N to 1 Task



2 Assignment 1 Discussion

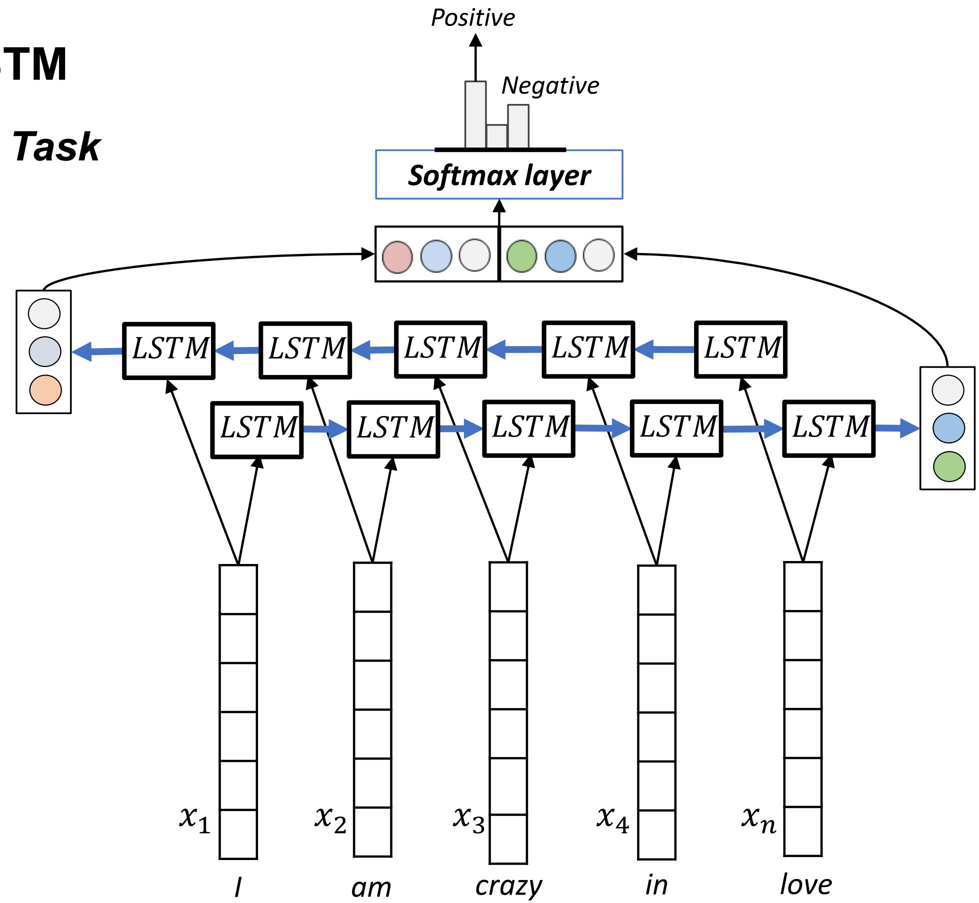
Bi-RNN

N to 1 Task



Bi-LSTM

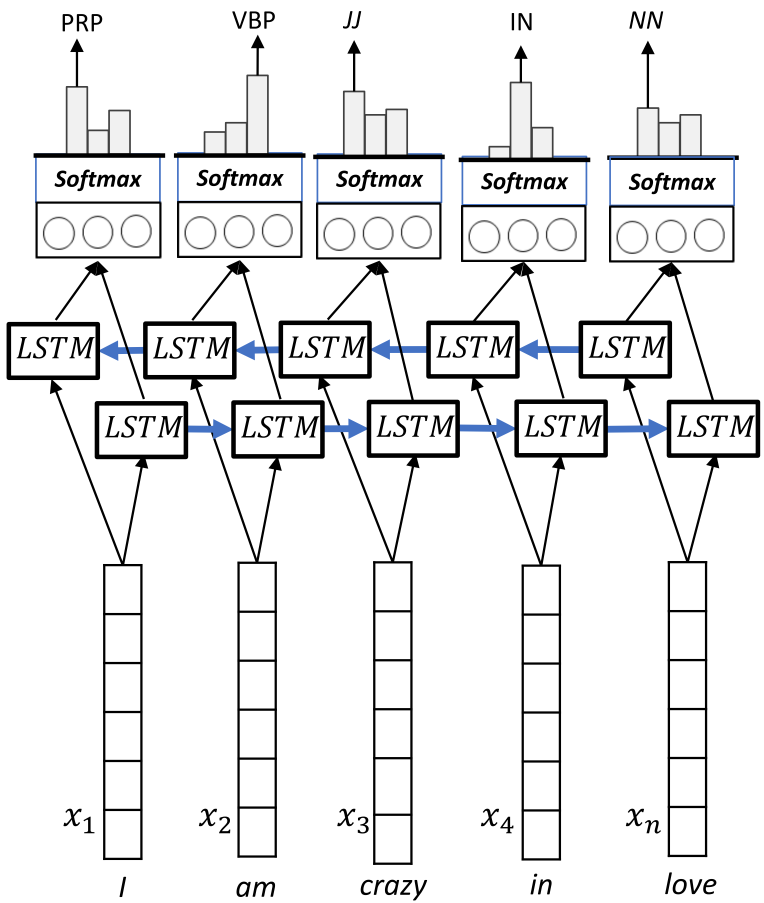
N to 1 Task



2 Assignment 1 Discussion

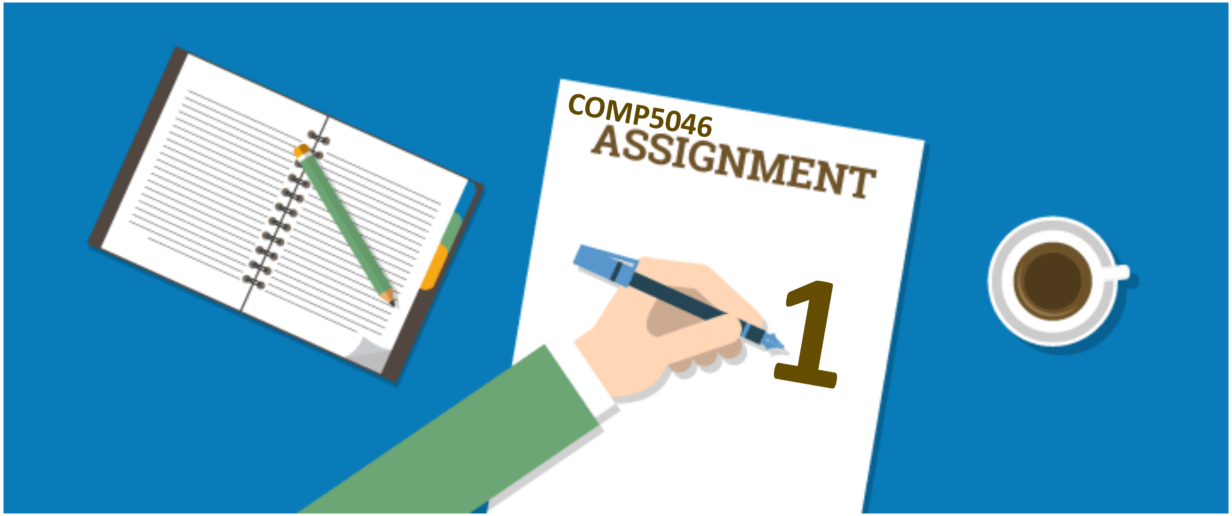
Bi-LSTM

N to N Task



2 Assignment 1 Discussion

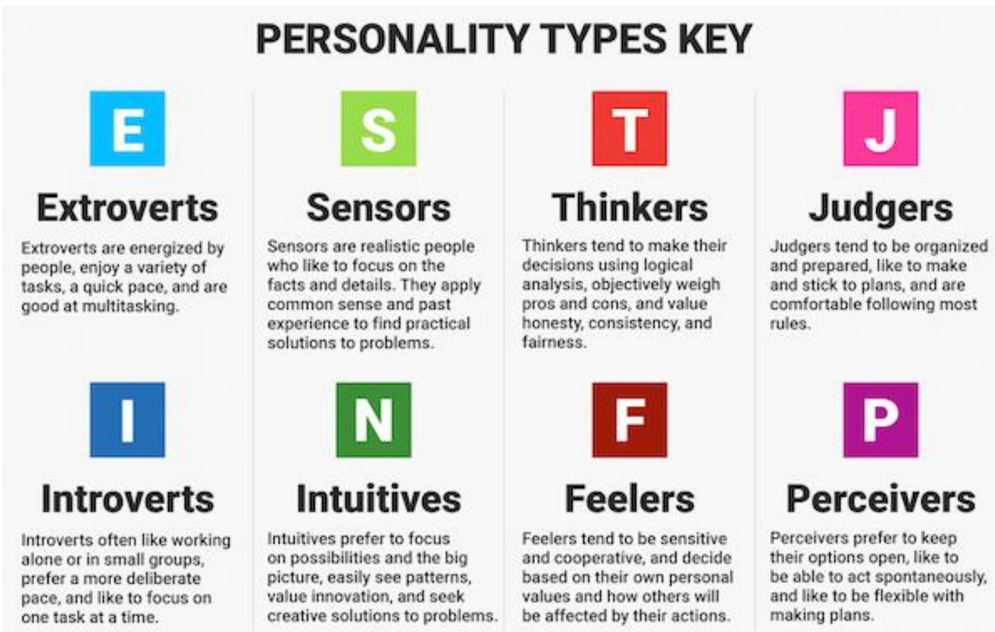
Let's discuss our Assignment 1



Assignment 1 Discussion

Understanding Personality:

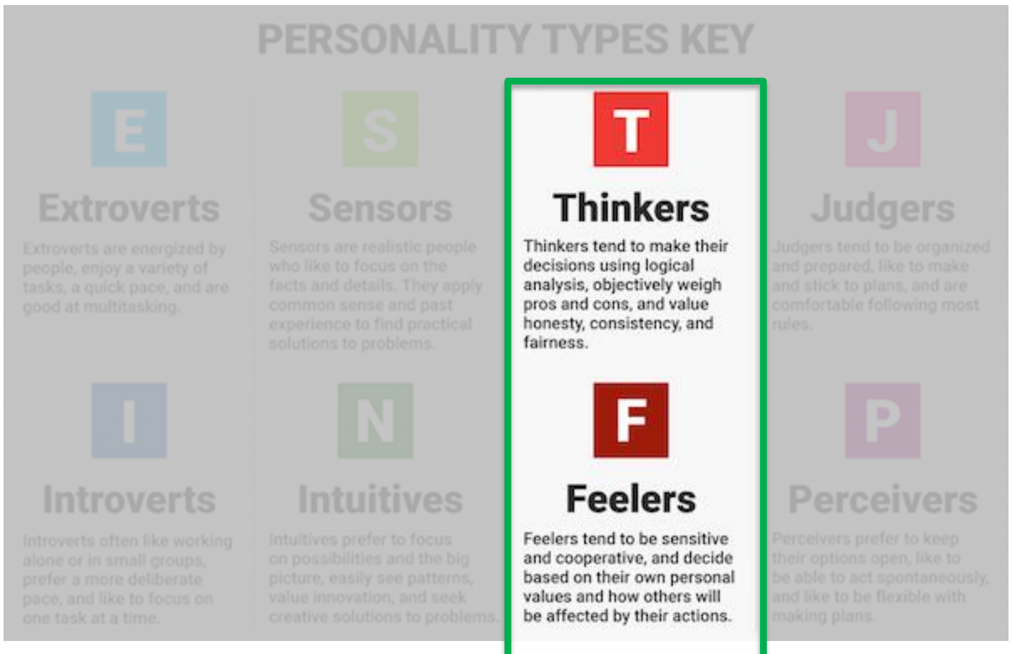
M**YERS – **B**RIGGS **T**YPE **I**NDICATOR **(MBTI)



2 Assignment 1 Discussion

Understanding Personality:

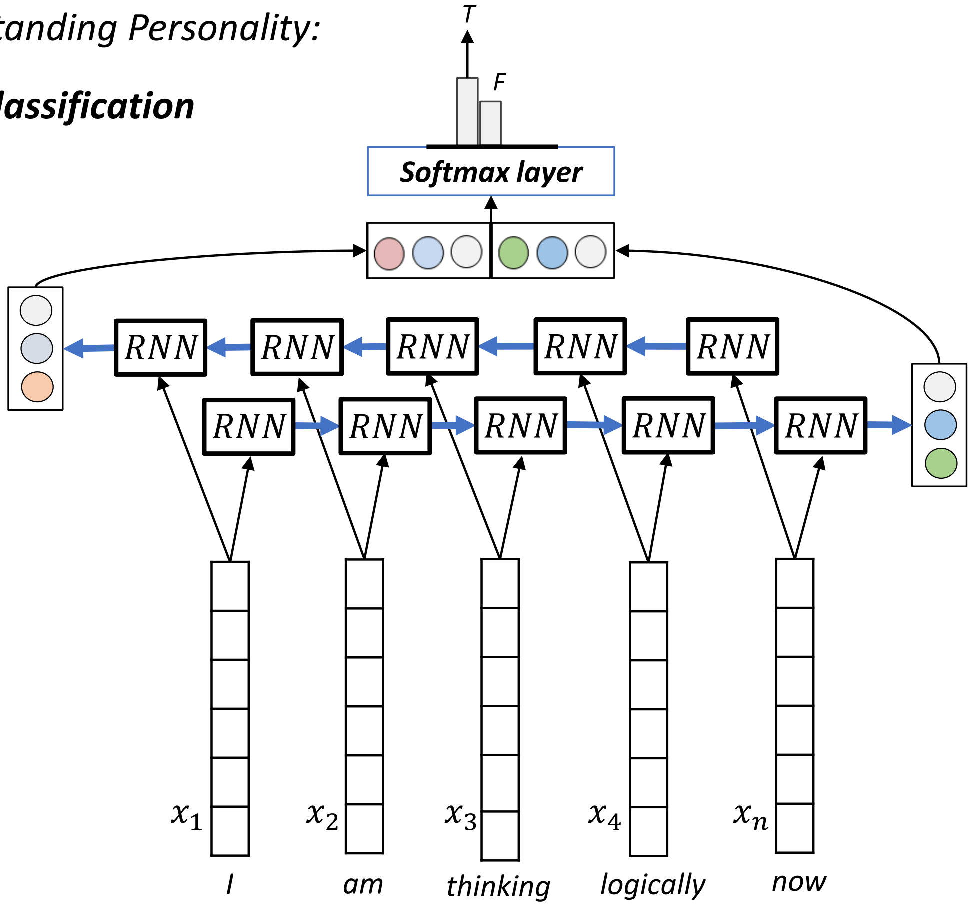
M**YERS – **B**RIGGS **T**YPE **I**NDICATOR **(MBTI)



2 Assignment 1 Discussion

Understanding Personality:

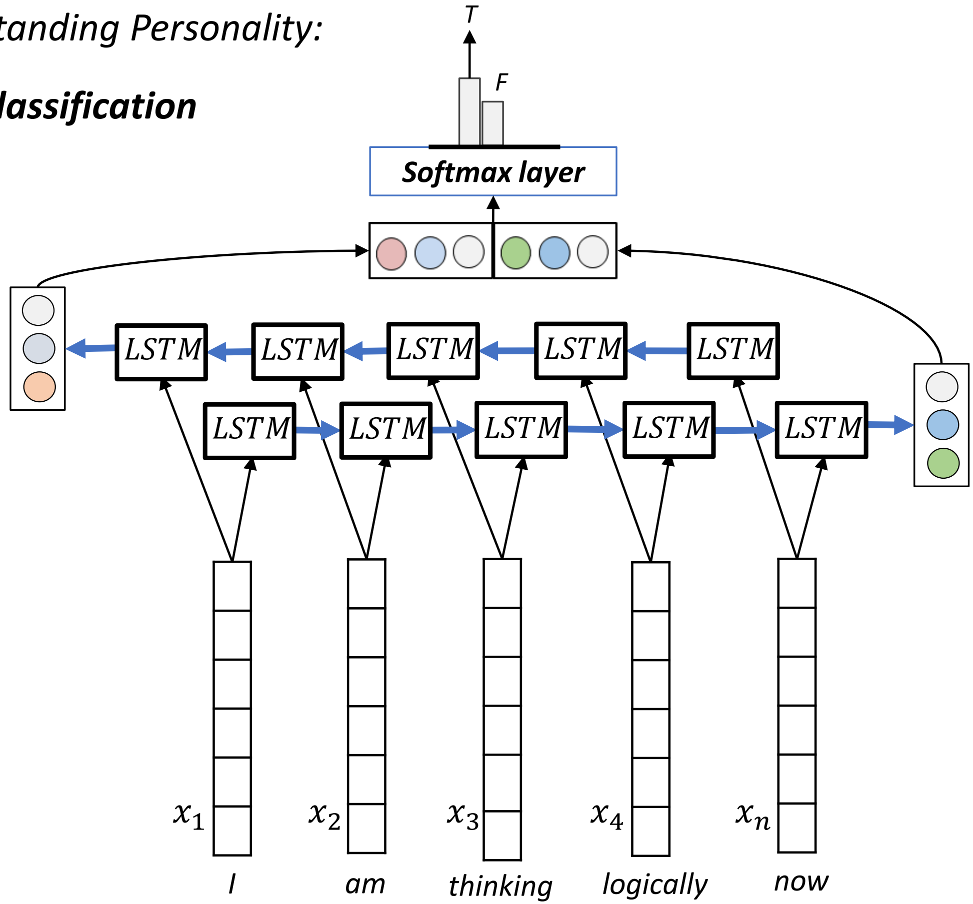
MBTI Classification



2 Assignment 1 Discussion

Understanding Personality:

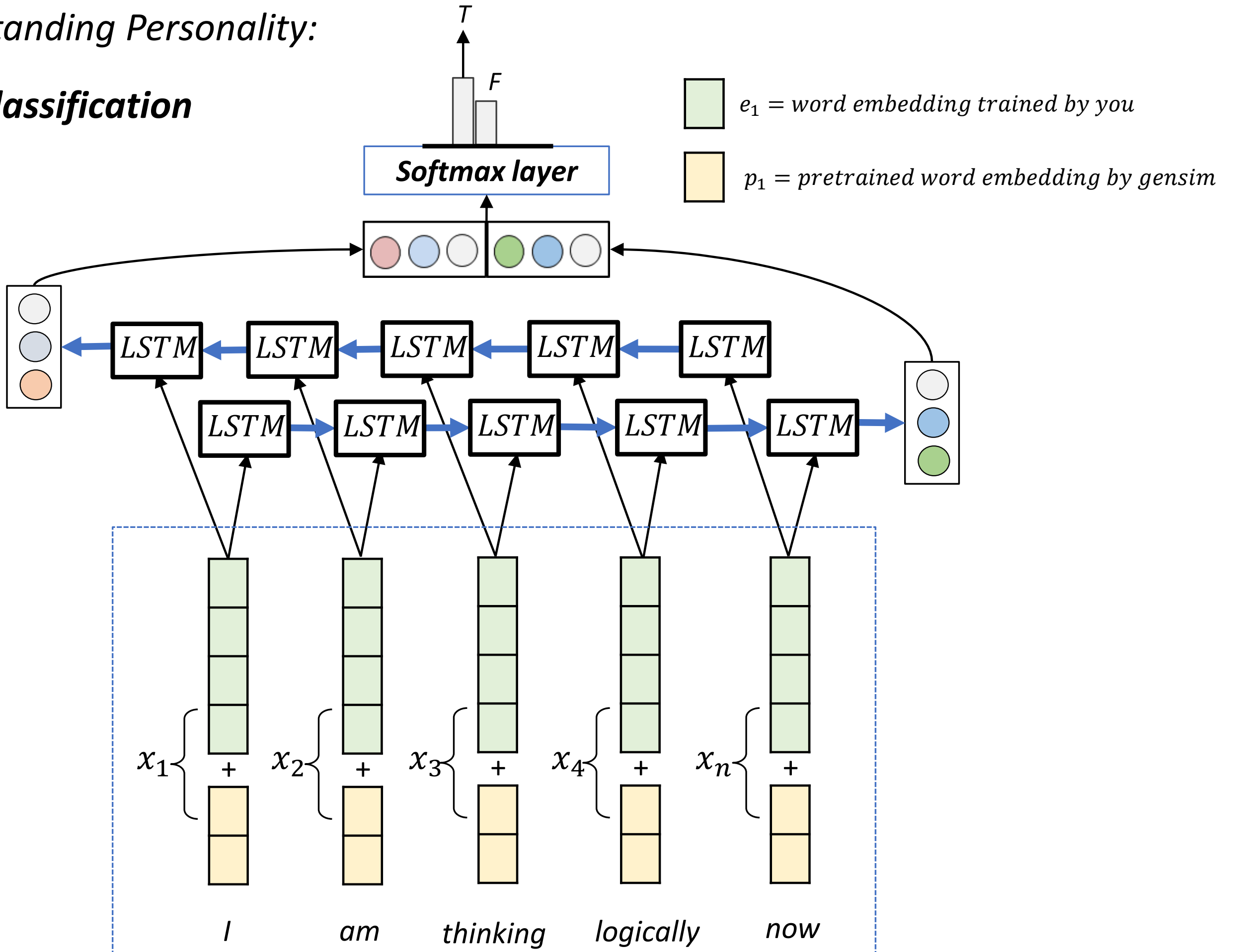
MBTI Classification



2 Assignment 1 Discussion

Understanding Personality:

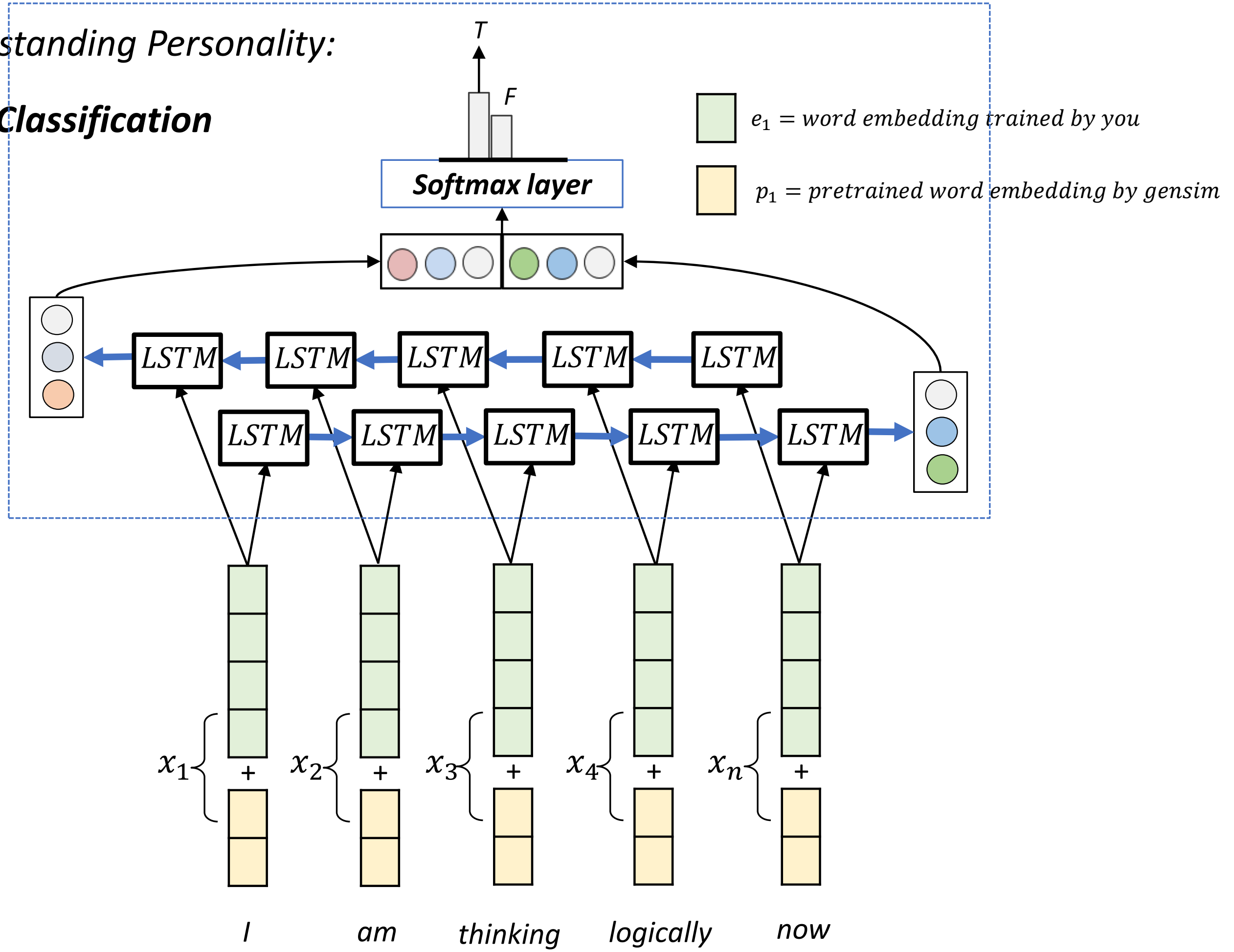
MBTI Classification



2 Assignment 1 Discussion

Understanding Personality:

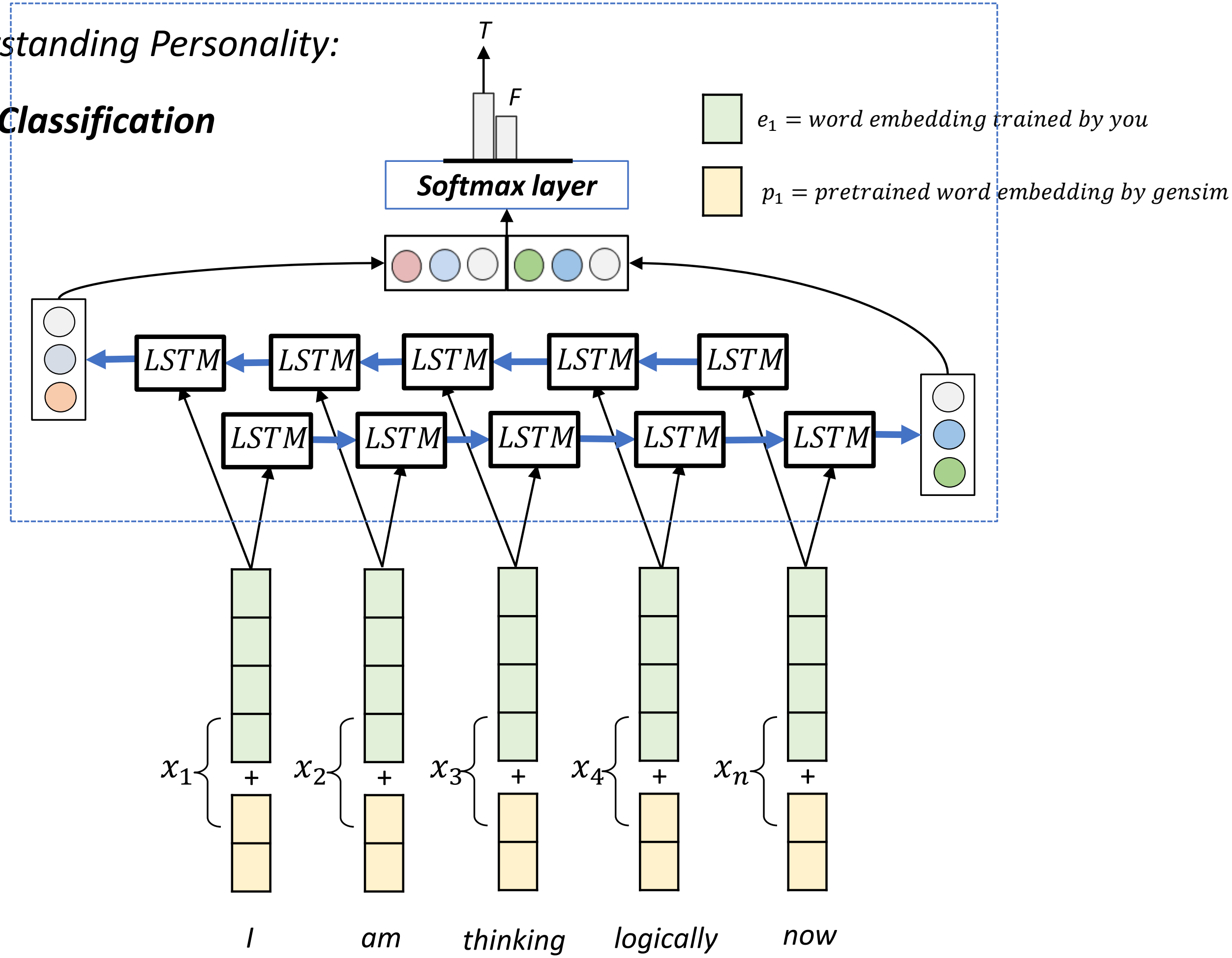
MBTI Classification



Assignment 1 Discussion

Understanding Personality:

MBTI Classification



Intrinsic word vector evaluation

Evaluation Result Comparison

The Semantic-Syntactic word relationship tests for understanding of a wide variety of relationships as shown below.

Table 2: Results on the word analogy task, given as percent accuracy. Underlined scores are best within groups of similarly-sized models; bold scores are best overall. HPCA vectors are publicly available[2]; (i)vLBL results are from (Mnih et al., 2013); skip-gram (SG) and CBOW results are from (Mikolov et al., 2013a,b); we trained SG† and CBOW† using the `word2vec` tool[3]. See text for details and a description of the SVD models.

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	<u>67.5</u>	<u>54.3</u>	<u>60.3</u>
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	<u>64.8</u>	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	<u>80.8</u>	61.5	<u>70.3</u>
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	<u>67.4</u>	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	<u>77.4</u>	67.0	<u>71.7</u>
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	<u>**81.9**</u>	<u>**69.3**</u>	<u>**75.0**</u>

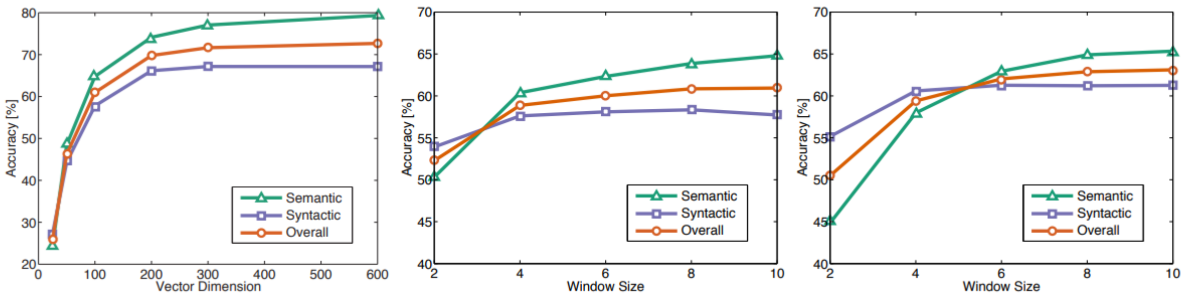
(Original Glove Paper - Pennington et al.2014)

Intrinsic word vector evaluation

Evaluation Result Comparison

The Semantic-Syntactic word relationship tests for understanding of a wide variety of relationships as shown below.

Window-Size (m) and Vector Dimension (N)

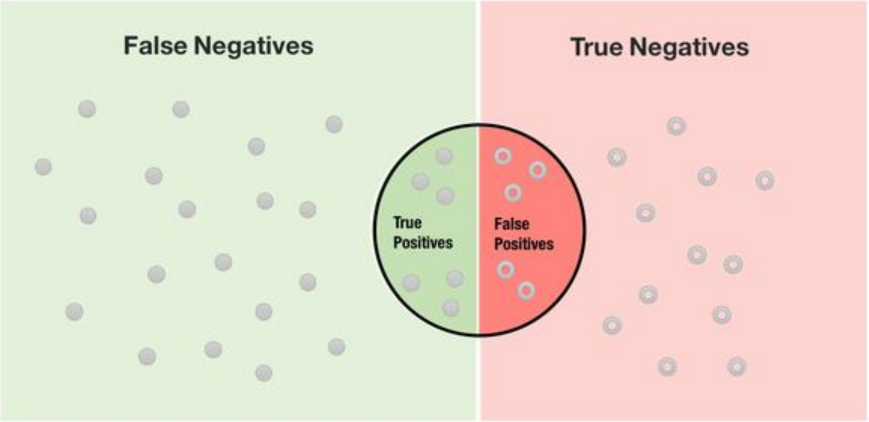


(a) Symmetric context (b) Symmetric context (c) Asymmetric context

Figure 2: Accuracy on the analogy task as function of vector size and window size/type. All models are trained on the 6 billion token corpus. In (a), the window size is 10. In (b) and (c), the vector size is 100.

(Original Glove Paper - Pennington et al.2014)

Assignment 1: Performance Evaluation



$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

Assignment 1 Discussion

Assignment 1: Performance Evaluation

Evaluation 2)

Model	F1
Bi-LSTM With URL	Xxx
Bi-LSTM Without URL	Xxx
Bi-LSTM With stopword	Xxx
Bi-LSTM Without stopword	Xxx
…	…

Evaluation 3)

Model	F1
Bi-LSTM with Word2vec (SG)	Xxx
Bi-LSTM with Word2vec (CBOW)	Xxx
Bi-LSTM with Word2vec (CBOW) + glove-twitter-100	Xxx
…	…

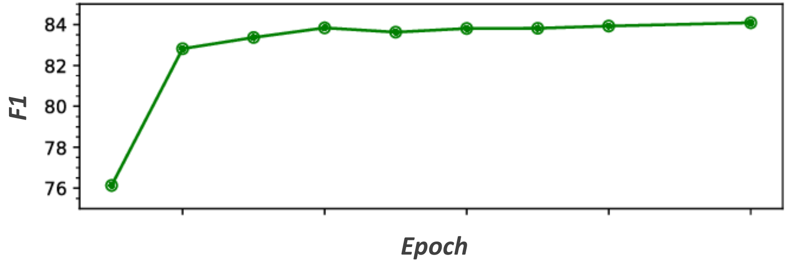
2 Assignment 1 Discussion

Assignment 1: Performance Evaluation

Evaluation 4)

Model	F1
Bi-RNN	Xxx
Bi-LSTM	Xxx
Bi-GRU	Xxx
…	…

Evaluation 5)



0 LECTURE PLAN

Lecture 5: Assignment1 and Language Fundamental

1. RNN/LSTM, Dealing Context Review
2. Assignment 1 Discussion
3. **Sentiment Analysis**
    1. Sentiment Analysis
    2. Sentiment Analysis: Examples
    3. Sentiment Analysis: Lexicons
4. Language Fundamental
    - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
    1. Tokenization
    2. Cleaning and Normalisation
    3. Stemming and Lemmatisation
    4. Stopword
    5. Regular Expression

The purpose of Natural Language Processing: Overview

Application

Understanding

Searching

Dialog

Translation

Sentiment Analysis

Topic Classification

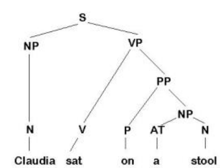
Search

Topic Modelling

....

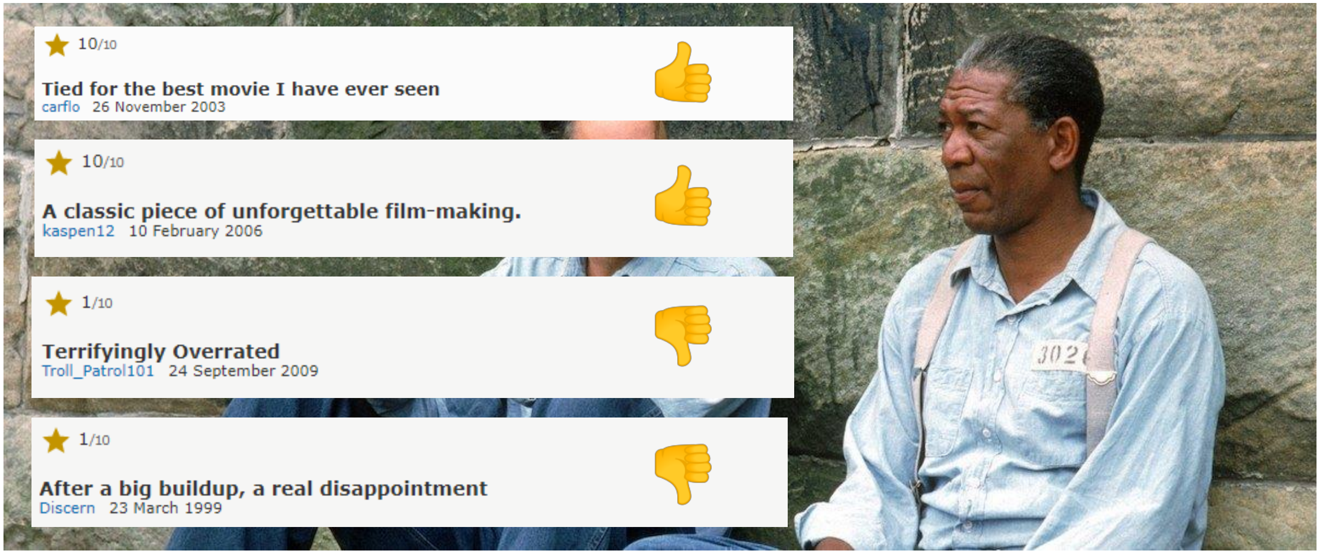
....

NLP Stack

Entity Extraction	When Sebastian Thrun …	When Sebastian Thrun PERSON started at Google ORG in 2007 DATE
Parsing	Claudia sat on a stool	
PoS Tagging	She sells seashells	**[she/PRP] [sells/VBZ] [seashells/NNS]**
Stemming	Drinking, Drank, Drunk	**Drink**
Tokenisation	How is the weather today	**[How] [is] [the] [weather] [today]**

3 Sentiment Analysis

Movie Review – Positive or Negative



Too easy? 😏

3 Sentiment Analysis

What is Sentiment Analysis?

3 Sentiment Analysis

What is Sentiment Analysis?

*"Sentiment analysis is the operation of **understanding the intent or emotion behind a given piece of text**. It is part of text classification, but it is useful for extracting structured information"*

Different Names of a 'Sentiment Analysis'

- *Opinion extraction*
- *Opinion mining*
- *Sentiment mining*
- *Subjectivity analysis*

Sentiment Analysis

Cottonelle FreshCare Flushable Wipes for Adults, Wet Wipes, Alcohol Free, 336 Wet Wipes per Pack (Eight 42-Count Resealable Soft Packs)

by Cottonelle

11,351 ratings

Available from these sellers.

Style Name: **8 Packs of Flushable Wipes**

- Superior Clean CleaningRipples texture provides softness while removing more cleans better versus using dry bath tissue alone
- 100 percent flushable & the No. 1 Flushable Wipe Brand among national flushable wipes brands
- Immediately Starts to Break Down After Flushing – Cottonelle bathroom wipes break down 6X's faster than Dude Wipes (based on strength loss testing) and are sewer safe & septic safe with SafeFlush Technology
- Moist wipes made from fibers that are 100 percent biodegradable
- Adult wipes that are infused with the gentle cleansing power of water and are perfect for man wipes, feminine wipes and more

Customer reviews

4.6 out of 5

11,351 customer ratings

5 star	79%
4 star	9%
3 star	5%
2 star	2%
1 star	4%

How does Amazon calculate star ratings?

Review this product

Share your thoughts with other customers

Write a customer review

Top international reviews

MustLoveDogs

Just because you CAN flush it, doesn't mean you should!

Reviewed in the United States on 14 July 2018

Style: 8 Packs of Flushable Wipes | **Verified Purchase**

Flushable? Not according to the plumber I just paid $200 to. Be careful folks. Other than the misleading "flushable" advertising, I liked product, but can't afford plumbing bills.

354 people found this helpful

Helpful | Report abuse

Zack Fischmann

These are NOT unscented -- one of the ingredients is "fragrance/parfum"

Reviewed in the United States on 15 January 2019

Style: 8 Packs of Flushable Wipes | **Verified Purchase**

3 Sentiment Analysis

What is Sentiment Analysis?

Emotion, Mood, Interpersonal stances, **Attitude**, Personality traits

Typology of Affective States (Scherer et al. 2006)

Attitudes

*Enduring, affectively colored beliefs, **dispositions towards objects/persons***

- *liking, loving, hating, valuing, desiring*

Scherer, K., Dan, E., & Flykt, A. (2006). What determines a feeling's position in affective space? A case for appraisal. Cognition & Emotion, 20(1), 92-113

Sentiment Analysis: Examples

Apple iPhone 7 - 128GB - Rose Gold (Unlocked)

★★★★½ 39 product ratings | About this product

Ratings and reviews

4.6

★★★★½

39 product ratings

★ 5		30
★ 4		7
★ 3		0
★ 2		1
★ 1		1

Aspects

92%	84%	51%
Good value	Long battery life	Small form factor

Write a review

Most relevant reviews

See all 24 reviews

★★★★★
by judee12
18 Jul, 2019

Excellent phone

Works excellently well, the screen is very very clear. Photos are better than my iPhone 5se, even though they are both 12mp. Front facing camera is 7mp, 5se is less. The only downside is the battery life. It doesn't last all day for me. I have small hands but the larger size isn't too big. Can highly recommend, good value.

Verified purchase: Yes | Condition: Pre-Owned

★★★★
by noadaughert_31
26 Apr, 2018

Really good for price

Had virtually no scratches and battery life is optimal despite being referbished. Good value for your money. Only complaint was that there wasnt any accessories such as the bluetooth ear buds required for listening to music or the lightning to AUX adapter. But no accessories were listed in the description.

Verified purchase: Yes | Condition: Pre-Owned

★★★★
by diannpedlo_0
03 Jan, 2019

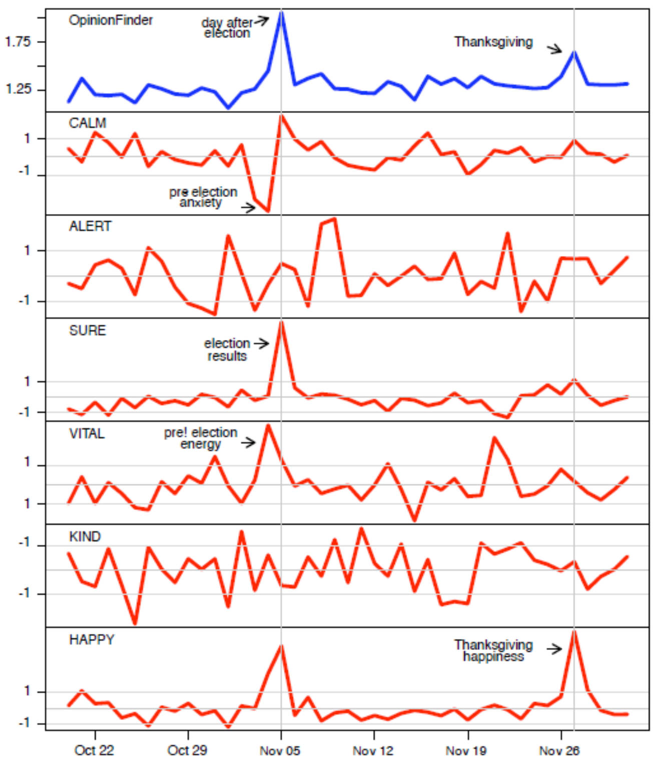
Good practical iPhone.

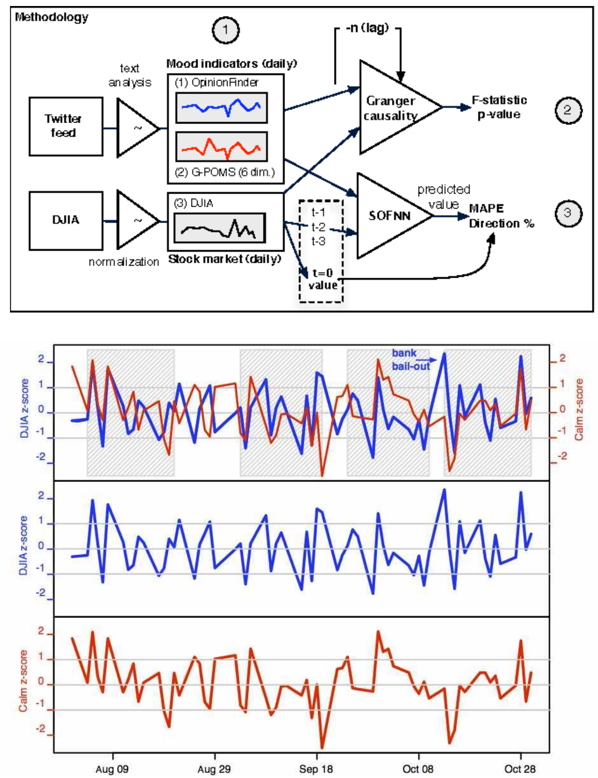
It's just so much better than my previous iPhone 6 as it was damaged & difficult to use. The iPhone 7 feels good to use. I'm not really sure it was the best price as I didn't shop around but am happy regardless,

Verified purchase: Yes | Condition: New

Sentiment Analysis: Examples

Twitter mood predicts the stock market (Bollen et al. 2011)





3 Sentiment Analysis

Sentiment Analysis: Sentiment viz

https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Sentiment Analysis Tasks

- ***Movie****: Is this review positive or negative?*
- ***Products****: what do people think about the new phone?*
- ***Public sentiment****: how is consumer confidence? Is despair increasing?*
- ***Politics****: what do people think about this candidate or issue?*
- ***Prediction****: predict election outcomes or market trends from sentiment*

Sentiment Analysis

What will be considered to analyse sentiment

Sentiment analysis = the detection of Attitudes

Enduring, affectively colored beliefs, dispositions towards objects/persons

Main Factors

- ***Target Object***: *an entity that can be a product, person, event, organisation, or topic (e.g. iPhone)*
- ***Attribute***: *an object usually has two types of attributes*
  - *Components (e.g. touch screen, battery)*
  - *Properties (e.g. size, weight, colour, voice quality)*
  - *Explicit and implicit attributes:*
    - *Explicit attributes: appearing in the attitude (e.g. "the battery life of this phone was not long")*
    - *Implicit attributes: not appearing in the attitude (e.g. "this phone is too expensive" – the property price)*
- ***Attitude Holder***: *the person or organisation that expresses the opinion (e.g. my mother was mad with me)*
- ***Type of attitude***: *positive, negative, or neutral or set of types (e.g. happy)*
- ***Time***: *the time that expresses the opinion*

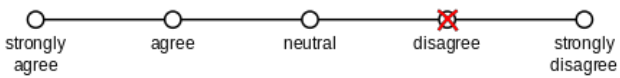
3 Sentiment Analysis

What is Sentiment Analysis?

- *Basic Task: Is the attitude of this text positive or negative?*

- *More complex task: Rank the attitude of this text from 1 to 5*

Likert Scale (1 to 5)



- *Advanced task: Detect the target, source, or complex* ***attitude types***

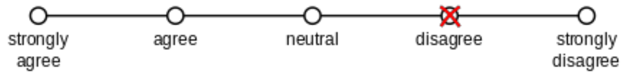
3 Sentiment Analysis

What is Sentiment Analysis?

- *Basic Task: Is the attitude of this text **positive or negative**?*

- *More complex task: Rank the attitude of this text from 1 to 5*

Likert Scale (1 to 5)



- *Advanced task: Detect the target, source, or complex attitude types*

3 Sentiment Analysis

Finding aspect/attribute/target of sentiment

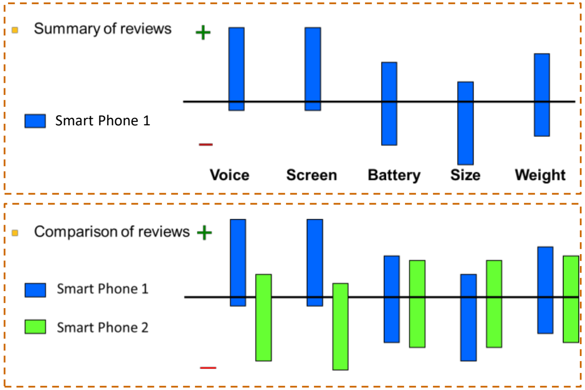
Title: Sharp, Solid, but Harder to Hold than IPhone 7
- By Tristan on March 13, 2017

"my thoughts on the iPhone 7 are:
1) Retina display is awesome. Everything looks more defined and sharper. There is much color and clarity out there... or should I say, in those digital images and videos... needless to say, the camera as well captures great images.
......."

Attribute based Summary

- Attribute 1: display
  - Positive
    1. Retina display is awesome
    2. There is much color and clarity out there
    3. …
- Attribute 2: camera
  - Positive
    1. the camera as well captures great images.
    2. .....

Attribute based Visualisation



Sentiment Analysis

Features Vectors: a bird's eye view

- Word ngrams (up to 4), skip ngrams w/ 1 missing word
- Character ngrams up to 5
- All caps: number of words in capitals
- Number of continuous punctuation marks, either exclamation or question or mixed. Also whether last char contains one of these.
- Presence of emoticons

Classify your Sentiment is a classification problem

- *Typically people have used **Naïve Bayes** or **Support Vector Machines (SVM)** in the past [Mohammad et al. 2013]*
- ***Artificial Neural Nets** are also becoming more popular now [Nogueira dos Santos & Gatti, 2014]*

Useful Sentiment Lexicons

Name	Details
The General Inquirer http://www.wjh.harvard.edu/~inquirer http://www.wjh.harvard.edu/~inquirer/homecat.htm http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls	Categories • Positive (1915 words) and Negative (2291 words) • Strong vs Weak, Active vs Passive, Overstated versus Understated • Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc Free to use
LIWC Linguistic Inquiry and Word Count http://www.liwc.net/	2300 words and less than 70 classes Affective Processes • negative emotion (bad, weird, hate, problem, tough) • positive emotion (love, nice, sweet) Cognitive Processes • Tentative (maybe, perhaps, guess), Inhibition (block, constraint) • Pronouns, Negation (no, never), Quantifiers (few, many) $30 or $90 fee
MPQA Subjectivity Cues Lexicon http://www.cs.pitt.edu/mpqa/subj_lexicon.html	Each word annotated for intensity (strong, weak) 6885 words from 8221 lemmas • 2718 positive • 4912 negative GNU GPL (widely-used free software license)
Opinion Lexicon http://www.cs.uic.edu/~liub/FBS/opinion--lexicon--English.rar	6786 words • 2006 positive/ 4783 negative Free to use
SentiWordNet http://swn.isti.cnr.it/	All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness • [estimable(J,3)] "may be computed or estimated" Pos 0 Neg 0 Obj 1 • [estimable(J,1)] "deserving of respect or high regard" Pos .75 Neg 0 Obj .25 Free to use

3 Sentiment Analysis

Can you build the sentiment lexicon by yourself?

Bootstrap style: Semi-supervised learning of lexicons

- *Use a small amount of information*
- *A few labeled examples*
- *A few hand--built patterns*
- *Bootstrapping a lexicon*

amazonmechanical turk
beta
Artificial Artificial Intelligence

0 LECTURE PLAN

Lecture 5: Assignment1 and Language Fundamental

1. RNN/LSTM, Dealing Context Review
2. Assignment 1 Discussion
3. Sentiment Analysis
   1. Sentiment Analysis
   2. Sentiment Analysis: Examples
   3. Sentiment Analysis: Lexicons
4. **Language Fundamental**
   - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
   1. Tokenization
   2. Cleaning and Normalisation
   3. Stemming and Lemmatisation
   4. Stopword
   5. Regular Expression

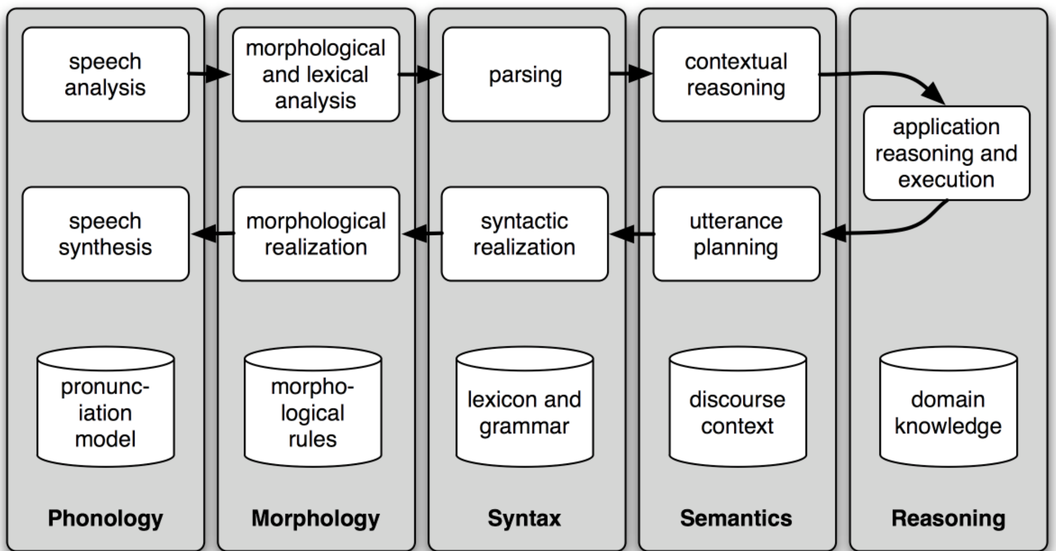
4 Language Fundamental

Level of Natural Language Processing

Level of Natural Language Processing

Phon

All so
system s



ragmatics

nguage use

4 Language Fundamental

We know the sounds of our language

Which sounds are in our language and which sounds are not

- For example, English speakers know the [ŋ] sound (in sing) does not appear at the beginning of a word
- Does this mean that [ŋ] cannot appear at the beginning of words in all human languages?

NO! — **Ng**uyen Tran

NO! — Andrew **Ng**

4 Language Fundamental

We know how sounds can combine

Often shown when a word from one language is borrowed into another:

- McDonalds — in English consonant clusters allowed ([mk] and [ldz])becomes...

マクドナルド	麦当劳	맥도날드
Makudonarudo	Màidāngláo	Maegdonaldeu

in other language — consonant clusters are not allowed

4 Language Fundamental

Morphology: Pieces of words

- A field of linguistics focused on the study of the ***forms and formation of words in a language***

- Words in a language consist of one element or elements of meaning which are ***morphemes***
  - ***Morphemes*** are the pieces of words: <u>bases, roots and affixes (pre-fix, suffix).</u>

Unacceptable

un

Prefix (a type of affix)
A group of letters placed ***<u>before</u>*** *the root word*

accept

Root word
the central morpheme, the key element

able

Suffixes (a type of affix)
A group of letters placed ***<u>after</u>*** *the root word*

Language Fundamental

Morphology: Pieces of words

- A field of linguistics focused on the study of the ***forms and formation of words in a language***

- Words in a language consist of one element or elements of meaning which are ***morphemes***
  - ***Morphemes*** are the pieces of words: <u>bases, roots and affixes.</u>

- walk walked walking walks walk walk -ed walk -ing walk -s

Language Fundamental

Natural Language Processing Level

- **Phonology/Morphology: the structure of words**
  - *Unusually* is composed of a prefix *un-*, a stem *usual*, and an affix *–ly*. *Learned* is *learn* plus the inflectional affix *–ed*
- **Syntax: the way words are used to form phrases**
  - It is part of English syntax that a determiner such as *the* will come before a noun, and also that determiners are obligatory with certain singular noun.
- Semantics: Compositional and lexical semantics
  - Compositional semantics: the construction of meaning based on syntax
  - Lexical semantics: the meaning of individual words
- Pragmatics: meaning in context
  - *Do you have the time?* – means *'can you tell me what time is it now?'*

0 LECTURE PLAN

Lecture 5: Assignment1 and Language Fundamental

1. RNN/LSTM, Dealing Context Review
2. Assignment 1 Discussion
3. Sentiment Analysis
    1. Sentiment Analysis
    2. Sentiment Analysis: Examples
    3. Sentiment Analysis: Lexicons
4. Language Fundamental
    - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. **Text Preprocessing**
    1. Tokenization
    2. Cleaning and Normalisation
    3. Stemming and Lemmatisation
    4. Stopword
    5. Regular Expression

Text Preprocessing

- Every NLP task needs to do text pre-processing
  - Segmenting/tokenizing words in running text
  - Normalizing word formats
  - Segmenting sentences in running text

5 Text Preprocessing

How many words?

- Type: an element of the vocabulary.
- Token: an instance of that type in running text.

- How many of them in the sentence?
  - 14 tokens
  - 13 types (or 12) (or 11?)

they lay back on the Sydney grass and looked at the stars and their

- **Token** = number of tokens
- **Type** = vocabulary = set of types
  - $|V|$ is the size of the vocabulary

5 Text Preprocessing

How many words?

- ***N*** = number of tokens
- ***V*** = vocabulary = set of types
  - |*V*| is the size of the vocabulary

	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

Tokenization: language issues

- **French**
  - L'ensemble → one token or two?
    - L ? L' ? Le ?
    - Want l'ensemble to match with un ensemble
      - Until 2003, Google cannot make this work

- **German noun compounds are not segmented**
  - ***Lebensversicherungsgesellschaftsangestellter***
  - 'life insurance company employee'
  - German information retrieval needs ***compound splitter***

Text Preprocessing

Tokenization: language issues

- Chinese has no spaces between words:
  - 悉尼大学位于澳大利亚悉尼
  - 悉尼大学 位于 澳大利亚 悉尼
  - University of Sydney is located in Sydney, Australia

- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats

フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

Katakana　Hiragana　Kanji　Romaji

5 Text Preprocessing

Tokenization: language issues

- Arabic (or Hebrew) is basically written right to left, but with certain items like numbers written left to right
- Words are separated, but letter forms within a word form complex ligatures

← → ← → ← **start**

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

- ‘Algeria achieved its independence in 1962 after 132 years of French occupation.’
- With Unicode, the order of characters in files matches the conceptual order, and the reversal of displayed characters is handled by the rendering system.

Normalization

- Need to “normalize” terms
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match U.S.A. and USA
- We implicitly define equivalence classes of terms
  - e.g., deleting periods in a term
- Alternative: asymmetric expansion:
  - Enter: window Search: window, windows
  - Enter: windows Search: Windows, windows, window
  - Enter: Windows Search: Windows
- Potentially more powerful, but less efficient

Case Folding

- Applications like IR: ***convert all letters to lower case***
  - Since users tend to use lower case
  - Possible exception: upper case in mid-sentence?
    - e.g., General Motors
    - Fed vs. fed
    - SAIL vs. sail
- For sentiment analysis, Machine Translation, Information extraction
  - Case is helpful (US versus us is important)

Lemmatization

- Reduce inflections or variant forms to ***base form***
  - am, are, is → be
  - car, cars, car's, cars’ → car
- *the boy's cars are different colors → the boy car be different color*

- Lemmatization: have to find correct dictionary headword form
  Machine translation
  - Spanish quiero (‘I want’), quieres (‘you want’) same lemma as querer ‘want’

Morphology

- Morphemes:
  - The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Bits and pieces that adhere to stems
  - Often with grammatical functions

Text Preprocessing

find the root

stem

may not be actual word

lemma actual word

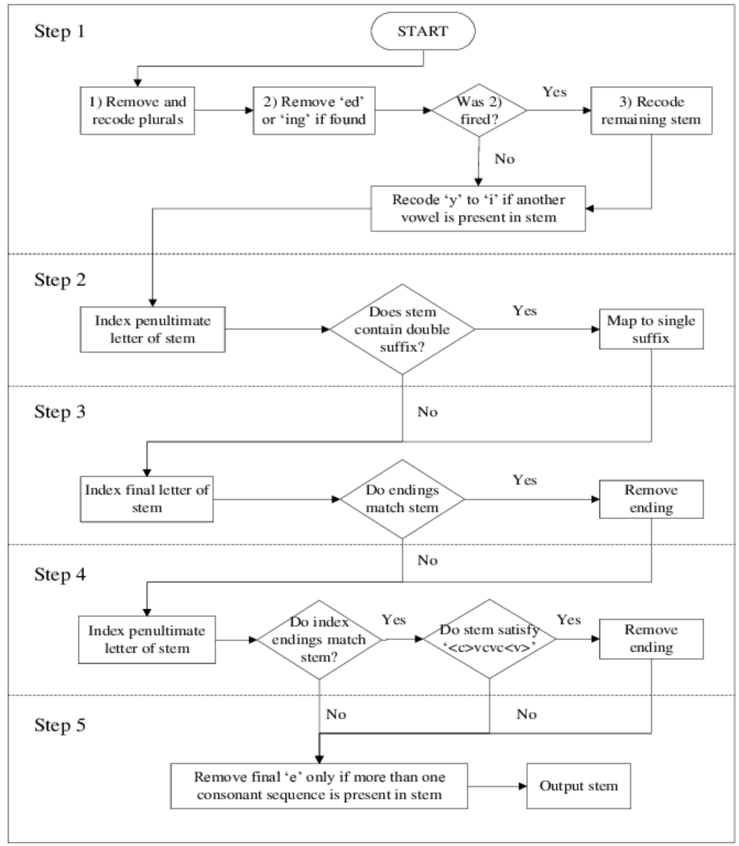
Stemming

- Reduce terms to their stems in information retrieval
- Stemming is crude chopping of affixes
  - language dependent
  - *e.g., automate(s), automatic, automation all reduced to automat.*

for example compressed and compression are both accepted as equivalent to compress.	→	for exampl compress and compress ar both accept as equival to compress

Porter's algorithm: The most common English stemmer

Porter Stemming Algorithm



5 Text Preprocessing

Dealing with complex morphology is sometimes necessary

- Some languages requires complex morpheme segmentation
  - Turkish
  - Uygarlastiramadiklarimizdanmissinizcasina
  - `(behaving) as if you are among those whom we could not civilize’
  - Uygar `civilized’ + las `become’
    - + tir `cause’ + ama `not able’
    - + dik `past’ + lar ‘plural’
    - + imiz ‘p1pl’ + dan ‘abl’
    - + mis ‘past’ + siniz ‘2pl’ + casina ‘as if’

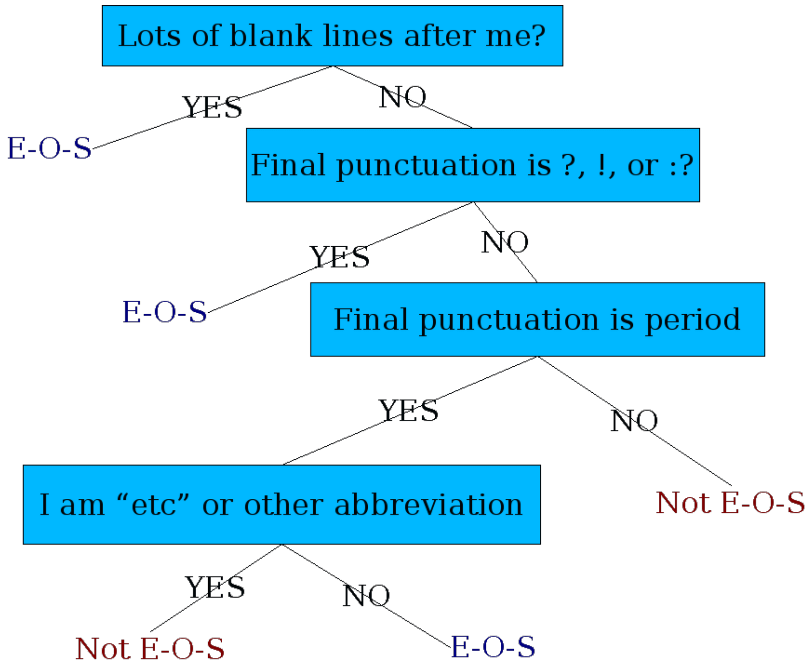
5 Text Preprocessing

Sentence Segmentation

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
  - Sentence boundary
  - Abbreviations like Inc. or Dr.
  - Numbers like .02% or 4.3
- Build a binary classifier
  - Looks at a "."
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning

5 Text Preprocessing

Sentence Segmentation using a Decision Tree



Implementing Decision Trees or other classifiers

- A decision tree is just an if-then-else statement
- The interesting research is choosing the features
- Setting up the structure is often too hard to do by hand
  - Hand-building only possible for very simple features, domains
    - For numeric features, it's too hard to pick each threshold
  - Instead, structure usually learned by machine learning from a training corpus

- As features that could be exploited by any kind of classifier
  - Logistic regression
  - SVM
  - Neural Nets
  - etc.

5 Text Preprocessing

Regular expressions

- A formal language for specifying text strings
- How can we search for any of these?
  1. woodchuck
  2. woodchucks
  3. Woodchuck
  4. Woodchucks

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges [A-Z]

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Regular Expressions: Negation in Disjunction

- Negations [^Ss]
  - Carat means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither ‘S’ nor ‘s’	I have no exquisite reason”
a^b	The pattern a carat b	Look up a^b now

5 Text Preprocessing

Regular Expressions: More Disjunction

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
`groundhog|woodchuck`	
`yours|mine`	`yours` `mine`
`a|b|c`	= [abc]
`[gG]roundhog|[Ww]oodchuck`	

5

Text Preprocessing

Regular Expressions: ?   * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun begun beg3n

Stephen C Kleene

Kleene *,   Kleene +

Regular Expressions: Anchors ^ $

Pattern	Matches
`^[A-Z]`	Palo Alto
`^[^A-Za-z]`	1 "Hello"
`\.$`	The end.
`.$`	The end? The end!

5 Text Preprocessing

Summary

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

/ Reference

Reference

- Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.