1. Introduction to NLP

Disclaimer

- Some slides are based on Stanford NLP/IR book (and slides)
 - https://nlp.stanford.edu/fsnlp/promo/
 - https://nlp.stanford.edu/IR-book/
- Some slides from Berkeley AI course
 - http://ai.berkeley.edu/home.html
- All three courses are available on Youtube (excellent follow-up materials of this course)

Introducing NLP Applications: Question Answering (QA)

Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S

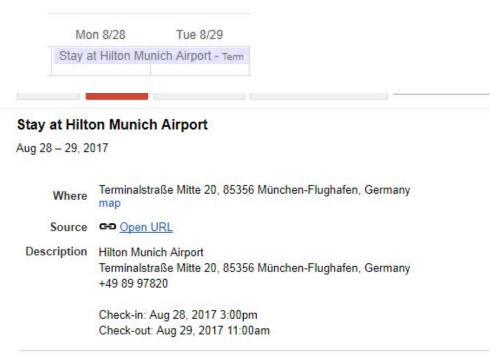
"AN ACCOUNT OF THE PRINCIPALITIES
OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL



Information Extraction (IE)







Dialogue (Limited)



Sentiment Analysis



Attributes:

zoom
affordability
size and weight
flash
ease of use

managery total harden. He wereast the has less upon people forced while the first and union. The eligible paper frames, in practs. I are priviley that, common with the Sergery carrie to trendity windle to if \$1, year among their Borney with, It has petitioned threat recommend many many care. Little book rands for a tigo postly, book falling system while it may to and other I have being the of tell. They make an arrival, and the industrial a fine for morne. If there are spray was a love a waste to copper plantics in yet or wonory's other bollocies one yet must be enough planting to the property of the same of the same of the same A mickly prior pourly (fight) we total among the authority to Lon Pyrou published with the smooth controlled that play control theory metaling such amind plants in you only a creat and how photographon. This only immaging thing is that it did from outside the appropriate spread presents out they seems CONTROL & MARKET CONTROL This oursers digital West Disease, it housing your, flads-group with EWay comgreat principle. I there you it me were provide notes. First it was difficult to with matrix, that argue language. Placely the Toward Prop. of proprior state in participants. Text. and real when I and removal etg. Board of the printed and security of the Livery or compatible recovery or \$15 Magnetis, \$1,435 per (0.56 Margoll), \$ warming a Queel and with the profession of the profession of the basis growing combons are a secretary by address and topography and supplied of accurate filled by atdistance of Armania, committee and your place have been alleged and Everyone equal time authors for second on hour. Here is what if hower to stay about it houses as mind that they only cause if for that large. Here demonstra all (to ret air print to the flow 2004 8020 bid steel proops For boy Purelia. Was now, I leaden of your hand your haid when guy had a

Size and weight



nice and compact to carry!



 since the camera is small and light, I around those heavy, bulky professio



the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

2. Rule-based Approaches

```
Classification:
```

```
F(text) is {spam,non-spam}
{positive, negative}
```

Let's build a rule-based spam filter

- "Viagra" or "Cialis"
- Is this a good rule?

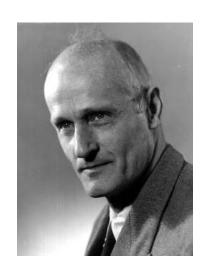
Regular expressions

- A formal language for specifying text strings
- How can we search for any of these?
 - woodchuck
 - woodchucks
 - Woodchuck 📮
 - Woodchucks



Regular Expressions: ? * +

| Pattern | Matches | |
|---------|----------------------------|----------------------------|
| colou?r | Optional previous char | <u>color</u> <u>colour</u> |
| oo*h! | 0 or more of previous char | oh! ooh! oooh! |
| o+h! | 1 or more of previous char | oh! ooh! oooh! |
| baa+ | | baa baaa baaaa |
| beg.n | | begin begun began |



Stephen C Kleene

Kleene *, Kleene +

Example

Find me all instances of the word "the" in a text.

```
the

Misses capitalized examples
```

[tT]he

Incorrectly returns other or theology

```
[^a-zA-Z][tT]he[^a-zA-Z]
```

How do we know if our rule is good?

- The process we just went through was based on fixing two kinds of errors
 - Matching strings that we should not have matched (there, then, other)
 - False positives (Type I)
 - Not matching things that we should have matched (The)
 - False negatives (Type II)



Precision vs. Recall

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
 - Increasing accuracy or precision (minimizing false positives)
 - Increasing coverage or recall (minimizing false negatives).



The 2-by-2 contingency table

| | correct | not correct | | | | | |
|--------------|---------|-------------|--|--|--|--|--|
| selected | tp | fp | | | | | |
| not selected | fn | tn | | | | | |

Precision and recall

• **Precision**: % of selected items that are correct

Recall: % of correct items that are selected



| | correct | not correct | | | | | |
|--------------|---------|-------------|--|--|--|--|--|
| selected | tp | fp | | | | | |
| not selected | fn | tn | | | | | |

A combined measure: F

 A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$



- The harmonic mean is a very conservative average; see https://machinelearningmastery.com/arithmetic-geometric-and-harmonic-means-for-machine-learning/
- People usually use balanced F1 measure

• i.e., with
$$\beta = 1$$
 (that is, $\alpha = \frac{1}{2}$): $F = \frac{2PR}{(P+R)}$

3. Word (Lexical) Similarity

Definition

How similar are two strings?

- Spell correction
 - The user typed "graffe"Which is closest?
 - graf
 - graft
 - grail
 - giraffe

- Computational Biology
 - Align two sequences of nucleotides

```
AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

Resulting alignment:

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

Also for Machine Translation, Information Extraction, Speech Recognition

Edit Distance

- The minimum edit distance between two strings
- Is the minimum number of editing operations
 - Insertion
 - Deletion
 - Substitution
- Needed to transform one into the other

Minimum Edit Distance

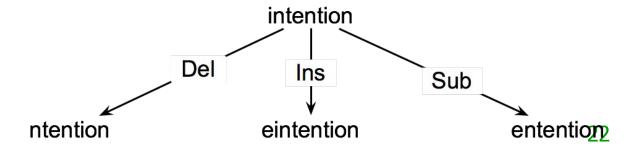
• Two strings and their **alignment**:

Minimum Edit Distance

- If each operation has cost of 1
 - Distance between these is 5
- If substitutions cost 2 (Levenshtein)
 - Distance between them is 8

How to find the Min Edit Distance?

- Searching for a path (sequence of edits) from the start string to the final string:
 - Initial state: the word we're transforming
 - Operators: insert, delete, substitute
 - Goal state: the word we're trying to get to
 - Path cost: what we want to minimize: the number of edits



Weighted Edit Distance

- Why would we add weights to the computation?
 - Spell Correction: some letters are more likely to be mistyped than others
 - Biology: certain kinds of deletions or insertions are more likely than others

Confusion matrix for spelling errors

| X | sub[X, Y] = Substitution of X (incorrect) for Y (correct) X Y (correct) | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|----|----|----|-----|---|----|----|-----|---|---|----|----|-----|----|----|---|----|----|----|----|---|-----|---|----|---|
| | a | ь | c | d | e | f | g | h | i | j | k | 1 | m | n | 0 | p | q | r | S | t | u | v | w | х | у | Z |
| a | 0 | 0 | 7 | 1 | 342 | 0 | 0 | 2 | 118 | 0 | 1 | 0 | 0 | 3 | 76 | 0 | 0 | 1 | 35 | 9 | 9 | 0 | 1 | 0 | 5 | 0 |
| b | 0 | 0 | 9 | 9 | 2 | 2 | 3 | 1 | 0 | 0 | 0 | 5 | 11 | 5 | 0 | 10 | 0 | 0 | 2 | 1 | 0 | 0 | 8 | 0 | 0 | 0 |
| c | 6 | 5 | 0 | 16 | 0 | 9 | 5 | 0 | 0 | 0 | 1 | 0 | 7 | 9 | 1 | 10 | 2 | 5 | 39 | 40 | 1 | 3 | 7 | 1 | 1 | 0 |
| d | 1 | 10 | 13 | 0 | 12 | 0 | 5 | 5 | 0 | 0 | 2 | 3 | 7 | 3 | 0 | 1 | 0 | 43 | 30 | 22 | 0 | 0 | 4 | 0 | 2 | 0 |
| С | 388 | 0 | 3 | 11 | 0 | 2 | 2 | 0 | 89 | 0 | 0 | 3 | 0 | 5 | 93 | 0 | 0 | 14 | 12 | 6 | 15 | 0 | 1 | 0 | 18 | 0 |
| f | 0 | 15 | 0 | 3 | 1 | 0 | 5 | 2 | 0 | 0 | 0 | 3 | 4 | 1 | 0 | 0 | 0 | 6 | 4 | 12 | 0 | 0 | 2 | 0 | 0 | 0 |
| g | 4 | 1 | 11 | 11 | 9 | 2 | 0 | 0 | 0 | 1 | 1 | 3 | 0 | 0 | 2 | 1 | 3 | 5 | 13 | 21 | 0 | 0 | 1 | 0 | 3 | 0 |
| h | 1 | 8 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 12 | 14 | 2 | 3 | 0 | 3 | 1 | 11 | 0 | 0 | 2 | 0 | 0 | 0 |
| i | 103 | 0 | 0 | 0 | 146 | 0 | 1 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 49 | 0 | 0 | 0 | 2 | 1 | 47 | 0 | 2 | 1 | 15 | 0 |
| j | 0 | 1 | 1 | 9 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| k | 1 | 2 | 8 | 4 | 1 | 1 | 2 | 5 | 0 | 0 | 0 | 0 | 5 | 0 | 2 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | . 4 | 0 | 0 | 3 |
| 1 | 2 | 10 | 1 | 4 | 0 | 4 | 5 | 6 | 13 | 0 | 1 | 0 | 0 | 14 | 2 | 5 | 0 | 11 | 10 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| m | 1 | 3 | 7 | 8 | 0 | 2 | 0 | 6 | 0 | 0 | 4 | 4 | 0 | 180 | 0 | 6 | 0 | 0 | 9 | 15 | 13 | 3 | 2 | 2 | 3 | 0 |
| n | 2 | 7 | 6 | 5 | 3 | 0 | 1 | 19 | 1 | 0 | 4 | 35 | 78 | 0 | 0 | 7 | 0 | 28 | 5 | 7 | 0 | 0 | 1 | 2 | 0 | 2 |
| 0 | 91 | 1 | 1 | 3 | 116 | 0 | 0 | 0 | 25 | 0 | 2 | 0 | 0 | 0 | 0 | 14 | 0 | 2 | 4 | 14 | 39 | 0 | 0 | 0 | 18 | 0 |
| p | 0 | 11 | 1 | 2 | 0 | 6 | 5 | 0 | 2 | 9 | 0 | 2 | 7 | 6 | 15 | 0 | 0 | 1 | 3 | 6 | 0 | 4 | 1 | 0 | 0 | 0 |
| q | 0 | 0 | 1 | 0 | 0 | 0 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| r | 0 | 14 | 0 | 30 | 12 | 2 | 2 | 8 | 2 | 0 | 5 | 8 | 4 | 20 | 1 | 14 | 0 | 0 | 12 | 22 | 4 | 0 | 0 | 1 | 0 | 0 |
| S | 11 | 8 | 27 | 33 | 35 | 4 | 0 | 1 | 0 | 1 | 0 | 27 | 0 | 6 | 1 | 7 | 0 | 14 | 0 | 15 | 0 | 0 | 5 | 3 | 20 | 1 |
| t | 3 | 4 | 9 | 42 | 7 | 5 | 19 | 5 | 0 | 1 | 0 | 14 | 9 | 5 | 5 | 6 | 0 | 11 | 37 | 0 | 0 | 2 | 19 | 0 | 7 | 6 |
| u | 20 | 0 | 0 | 0 | 44 | 0 | 0 | 0 | 64 | 0 | 0 | 0 | 0 | 2 | 43 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 8 | 0 |
| v | 0 | 0 | 7 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 8 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| w | 2 | 2 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 7 | 0 | 6 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| x | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| У | 0 | 0 | 2 | 0 | 15 | 0 | 1 | 7 | 15 | 0 | 0 | 0 | 2 | 0 | 6 | 1 | 0 | 7 | 36 | 8 | 5 | 0 | 0 | 1 | 0 | 0 |
| | Δ | Λ | Ω | 7 | Λ | Λ | Λ | 0 | Λ | Λ | Λ | 7 | 5 | Λ | Λ | n | 0 | 2 | 21 | 2 | Λ | Ω | Λ | Λ | 2 | Λ |

Keyboard used contributes significantly to model

