Photo by Raphael Schaller on Unsplash

COMP 3225

Natural Language Processing
2. Words

Prof Leslie Carr lac@soton.ac.uk
University of Southampton

In This Lecture You Will …

- Learn the basics of understanding language, by decomposing text into words
- Learn how to perform **basic text normalization tasks** including word segmentation and normalization, sentence segmentation, and stemming.
- Learn how to partition a text into tokens effectively using the BPE algorithm
- Learn how to create a minimum edit distance algorithm for comparing strings.
- Learn how to analyse the frequency of words in a text

The material in this lecture is based on chapter 2 of Jurafsky and Martin, Speech and Language Processing, 3rd edition (online) pp. 13–27

Understanding Computer Language

Lack of ambiguity

Clarity of meaning

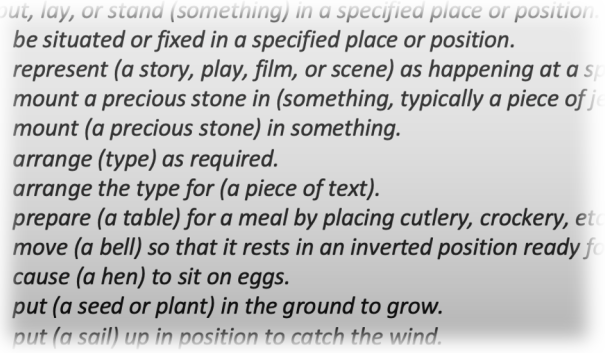
Allowed by programming techniques such as

- scope
- modularity
- closures
- API

vs Understanding Natural Language

- Ambiguity of identical word forms
  - Time flies like an arrow
    verb preposition
  - Fruit flies like a banana
    noun verb

The word with the most meanings in English is the word `set`*, with 430 senses listed in the OED.*



- Dependency on punctuation or intonation
  - James while John had had had had had had had had had had had a better effect on the teacher

Side note: understanding humour

- Humour often depends on ambiguity. The punchline reveals a second meaning.

- How do you make a Swiss roll?
  - Push him down the mountainside!
- How do you make a turtle fast?
  - Take away his food!

noun or verb?

verb or adjective?

- Double entendre
  - Panda mating fails: veterinarian takes over
  - New obesity study looks for larger test group
  - Children make nutritious snacks

Where to Start?

- In this course we'll look at understanding texts by looking at how words refer to named phenomena, how words combine to form complex concepts, how to model concepts as multidimensional vectors in abstract spaces. But we'll start off with words

"Words" by the Bee Gees (1967)

It’s Only Words: formal news

£500 to self-isolate under plan to slow Covid

Ministers hope cash will encourage self-isolation amid fears lockdown could last until summer

Chris Smyth, Whitehall Editor | Steven Swinford, Deputy Political Editor | George Grylls
Friday January 22 2021, 12.01am, The Times

Paying everyone who tests positive for coronavirus could cost £2 billion a month

Everyone in England who tests positive for coronavirus could be given £500 to ensure they self-isolate under plans to stop hardship spreading the virus.

Ministers are trying to solve a problem that scientific advisers have long said is an obstacle to controlling the virus. Paying all those with a positive test could cost £2 billion a month. However, the payment could be limited to those who cannot work from home.

The Times (newspaper)

It’s Only Words: polemic news

UP YOURS DELORS

At midday tomorrow Sun readers are urged to tell the French fool where to stuff his ECU

THE Sun today calls on its patriotic family of readers to tell the feelthy French to FROG OFF!

They **INSULT** us, **BURN** our lambs, **FLOOD** our country with dodgy food and **PLOT** to abolish the dear old pound.

Now it's your turn to kick **THEM** *in the Gauls.*

We want you to tell Froggie Common Market chief Jacques Delors exactly what you think of him and his countrymen.

By NICK PARKER, PETE WALSH and LIZ DUXBURY (Sun Diplomatic Staff)

At the stroke of noon tomorrow, we invite all true blue Brits to face France and yell "Up Yours, Delors."

The ear-bashing from our millions of readers will wake the EC President up to the fact that he will **NEVER** run our country.

His bid to replace the £ with the faceless ECU is the last straw after centuries of Froggy Brit-baiting. They:

BURNED alive British lambs earlier this year because they couldn't match our quality;

JEERED Mrs Thatcher when she visited Paris to boost celebrations for the bi-centenary of the French Revolution last year;

FOUL

BANNED British beef after falsely claiming it had mad cow disease;

BLEATED when we found their foul soft cheese was riddled with listeria bugs;

GAVE IN to the Nazis during the Second World War when we stood firm;

TRIED to conquer Europe until we put down Napoleon at Waterloo in 1815.

Remember, folks, it won't be long before the garlic-breathed bastilles will be here in droves once the Channel Tunnel is open.

So grab your megaphones, turn south and let 'em hear the British lion **ROAR**.

And the best of British to you all!

Where to bowl at the Gauls — Pages 2 and 3

The Sun (newspaper)

It’s Only Words: speech

Where 'd you try ?

Down the road I think .

Sainsbury 's .

You might

I did n't think you 'll mind having scrambled eggs with toast anyway .

Lovey dovey , .

You get more eggs this way , you get two eggs this way .

preferred the other way .

Not that .

Yeah , but you 're only giving us the.

If you always feel hungry what is the point ?

What diet ?

She 's lost two and a half stone within about two or three months .

Ha , what diet 's she on ?

Yeah , she works on this special diet plan , special diet and what have you .

I said well how do you manage to keep the weight off ?

Well she said I 've done a bit .

Why no I 've never seen you two

She is , she is jealous at having a dig at her .

No, no, I wo n't, I 'm, I 'm

Pretty cold .

If, if she had believed the aim to keep it off, I would 've been jealous , but she , she 's not going to find it easy to keep her weight, she 's out here when she 's put back on pound .

She 's tried .

What says you 're .

Yeah so am I .

Pardon ?

So am I what I 've left off .

Well I do n't know .

It 's all that work you do , cut the lawn .

Spoken conversation (BNC)

It’s Only Words: historic, poetic

To be, or not to be, that is the question:
Whether ’tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die—to sleep,
No more; and by a sleep to say we end
The heart-ache, and the thousand natural shocks
That flesh is heir to: ’tis a consummation
Devoutly to be wish’d. To die, to sleep.
To sleep, perchance to dream—ay, there’s the rub,
For in that sleep of death what dreams may come,
When we have shuffled off this mortal coil,
Must give us pause. There’s the respect
That makes calamity of so long life.
For who would bear the whips and scorns of time,
The oppressor’s wrong, the proud man’s contumely,
The pangs of dispriz’d love, the law’s delay,
The insolence of office, and the spurns
That patient merit of the unworthy takes,
When he himself might his quietus make
With a bare bodkin? Who would these fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The undiscover’d country, from whose bourn
No traveller returns, puzzles the will,
And makes us rather bear those ills we have
Than fly to others that we know not of?
Thus conscience does make cowards of us all,
And thus the native hue of resolution
Is sicklied o’er with the pale cast of thought,
And enterprises of great pith and moment,
With this regard their currents turn awry
And lose the name of action. Soft you now,
The fair Ophelia! Nymph, in thy orisons
Be all my sins remember’d.

Shakespeare

It’s Only Words: music lyrics

My anaconda don't
My anaconda don't
My anaconda don't want none
Unless you got buns, hun
Boy toy named Troy, used to live in Detroit
Big dope dealer money, he was gettin' some coins
Was in shootouts with the law, but he live in a palace
Bought me Alexander McQueen, he was keeping me stylish
Now that's real, real, real
Gun in my purse, bitch, I came dressed to kill
Who wanna go first? I had them pushin' daffodils
I'm high as hell, I only took a half a pill
I'm on some dumb shit
By the way, what he say?
He can tell I ain't missing no meals
Come through and him in my automobile
Let him eat it with his grills he keep tellin' me to chill
He keep telling me it's real, that he love my sex appeal
He say he don't like em boney, he want something he can grab
So I pulled up in the Jag and I hit him with the jab like
Dun-d-d-dun-dun-d-d-dun-dun
My anaconda don't
My anaconda don't
My anaconda don't want none
Unless you got buns, hun
Oh my gosh, look at her butt
Oh my gosh, look at her butt
Oh my gosh, look at her butt (Look at her butt)
Look at, look at, look at, look, at her butt

Clark Ernest Jr / Jones Jamal F

It’s Only Words: social media

WHO urges countries to 'track and trace' every Covid-19 case | World news | The Guardian\n\nMeanwhile the lump culls the population to fit his denuded NHS. https://t.co/lV6igoBDxl
@SepsisUK UK no longer tracking and tracing those suspected of having Coronavirus, going against \n@WHO advice (track and trace vital to isolate those infected to flatten curve of peak). UK strategy seems huge gamble, for highest risk groups and NHS capacity, without protective measures.
@fcbsd @chipps_sci @tombennett71 Track and trace isn't inaction\nHandwashing and hygiene campaigns aren't inaction.\nTelling all new symptomatic people to isolate isn't inaction.\nBriefing, tracing and preparation measures in the NHS are not inaction.
@RicHolden Agreed! But why stop mass testing? Track and trace will give true figures on demand placed upon NHS
15 #COVID_19uk\n#TeamCOVID19🖤\n\nThe World Health Organisation @WHO have said we must DETECT, it's a key part of their strategy/advice, to track and trace. \n\nThe NHS are NOT following this.\n\n#TeamNHS💙\n#TeamPatient💚 https://t.co/DJG1QgqPSv
Hallo @HelloFreshNL , ik heb geen track and trace ontvangen, box is niet geleverd en de huidige week staat niet meer in dr app?!
...Essentially electronic version of France’s paper forms. Use an app to say you are going out, have location tracking while out, and all this stored in a database. Data protected and only available to health service for retrospective track and trace...
@mancrepublic @Lambykins60 @omid9 UK no longer tracing and testing re contacts and Coronavirus goes against \n@WHO advice (track and trace vital to isolating infected - to flatten curve of peak - slow speed of community spread). NHS staff don’t have enough PPE to protect or adequate ICU capacity to save more lives
@OGiannino non so se possa funzionare: per superare la questione privacy non potremmo semplicemente avere una app da installare per il "track and trace" per coronavirus? Su base volontaria. Governativa. Chi non lo farebbe? Disinstallabile a piacimento. Se anche il 99% aderisse..
@itvnews @BorisJohnson Bro you aren't even testing NHS workers! TEST, TRACK AND TRACE

Twitter on Track & Trace 13/03/2020

Words about Words

Sentence	Unit of written language
Utterance	Unit of spoken language
Word Form	The inflected form as it actually appears in the corpus *e.g.* “said”
Lemma	An abstract form, shared by word forms having the same stem, part of speech, word sense *e.g.* “say” Stands for the class of words with same stem
Function words	Indicate the grammatical relationship between terms but have little topical information *e.g.* “by”
Types	Number of distinct words in a corpus *i.e.* vocabulary size
Tokens	Total collection of all words

Crude Lexical Analysis Pipeline

1. Original (Peter Pan at Project Gutenberg)

All children, except one, grow up. They soon know that they will grow up, and the way Wendy knew was this. One day when she was two years old, she was playing in a garden, and she plucked another flower and ran with it to her mother. I suppose she must have looked rather delightful, for Mrs. Darling put her hand to her heart and cried, “Oh, why can't you remain like this for ever!” This was all that passed between them on the subject, but henceforth Wendy knew that she must grow up. You always know after you are two. Two is the beginning of the end.

2. Simplified (punctuation stripped, case folded)

all children except one grow up they soon know that they will grow up and the way wendy knew was this one day when she was two years old she was playing in a garden and she plucked another flower and ran with it to her mother i suppose she must have looked rather delightful for mrs darling put her hand to her heart and cried oh why cant you remain like this for ever this was all that passed between them on the subject but henceforth wendy knew that she must grow up you always know after you are two two is the beginning of the end

3. Filtered (function words removed)

children except grow grow wendy knew years old playing garden plucked another flower ran mother suppose looked rather delightful mrs darling hand heart cried oh remain like passed subject henceforth wendy knew grow always beginning

4. Normalised (lemmatised and stemmed)

child except grow grow wendy know year old play garden pluck another flower run mother suppose look rather delight mrs darling hand heart cry oh remain like pass subject henceforth wendy know grow always begin

5. Index

always	1
another	1
begin	1
child	1
cry	1
darling	1
delight	1
except	1
flower	1
garden	1
grow	3
hand	1
heart	1
henceforth	1
know	2
like	1
look	1
mother	1
mrs	1
oh	1
old	1
pass	1
play	1
pluck	1
rather	1
remain	1
run	1
subject	1
suppose	1
wendy	2
year	1

Crude Bash Analysis Pipeline

1. Original (Peter Pan)

All children, except one, grow up. They soon know that they will grow up, and the way Wendy knew was this. One day when she was two years old, she was playing in a garden, and she plucked another flower and ran with it to her mother. I suppose she must have looked rather delightful, for Mrs. Darling put her hand to her heart and cried, "Oh, why can't you remain like this for ever!" This was all that passed between them on the subject, but henceforth Wendy knew that she must grow up. You always know after you are two. Two is the beginning of the end.

cat peter.txt

2. Simplified (punctuation stripped, case folded)

all children except one grow up they soon know that they will grow up and the way wendy knew was this one day when she was two years old she was playing in a garden and she plucked another flower and ran with it to her mother I suppose she must have looked rather delightful for mrs darling put her hand to her heart and cried oh why cant you remain like this for ever this was all that passed between them on the subject but henceforth wendy knew that she must grow up you always know after you are two two is the beginning of the end

tr A-Z a-z | tr –cs a-z '\n'

3. Filtered (function words removed)

children except grow grow wendy knew years old playing garden plucked another flower ran mother suppose looked rather delightful mrs darling hand heart cry oh remain like pass subject henceforth wendy knew grow always beginning

join –v 1 – STOPWORDS.txt. *#see next page*

4. Normalised (lemmatised / stemmed)

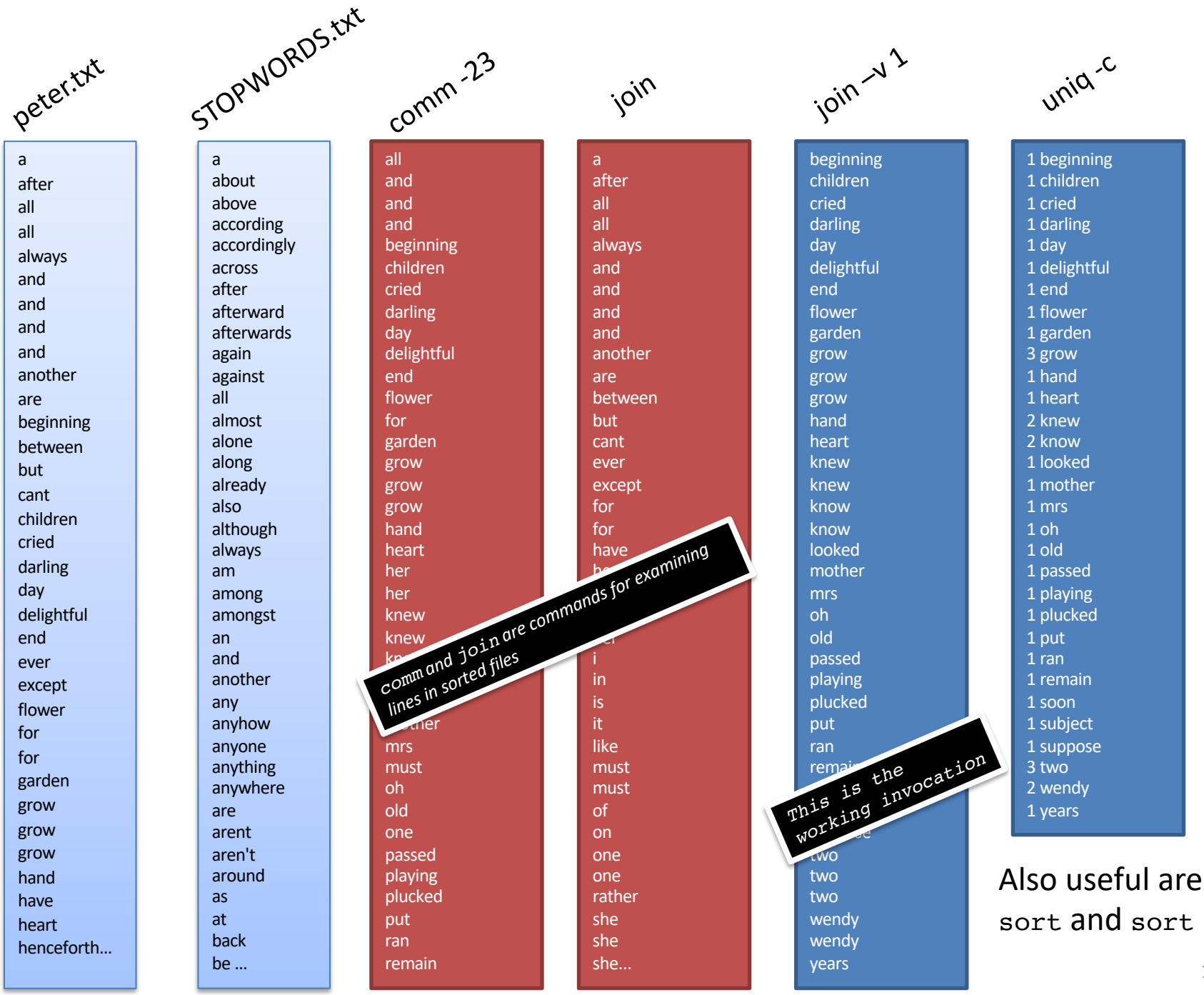
child except grow grow wendy know year old play garden pluck another flower run mother suppose look rather delight mrs darling hand heart cry oh remain like pass subject henceforth wendy know grow always begin

 # …see later…

5. Index

always	1
another	1
begin	1
child	1
cry	1
darling	1
delight	1
except	1
flower	1
garden	1
grow	3
hand	1
hear	1
henceforth	1
know	2
like	1
look	1
mother	1
mrs	1
oh	1
old	1
pass	1
play	1
pluck	1
rather	1
remain	1
run	1
subject	1
suppose	1
wendy	2
year	1

uniq –c


peter.txt
STOPWORDS.txt
comm -23
join
join –v 1
uniq -c
a after all all always and and and and another are beginning between but cant children cried darling day delightful end ever except flower for for garden grow grow grow hand have heart henceforth...
a about above according accordingly across after afterward afterwards again against all almost alone along already also although always am among amongst an and another any anyhow anyone anything anywhere are arent aren't around as at back be ...
all and and and beginning children cried darling day delightful end flower for garden grow grow grow hand heart her her knew knew mrs must oh old one passed playing plucked put ran remain
a after all all always and and and and another are between but cant ever except for for have i in is it like must must of on one one rather she she she...
beginning children cried darling day delightful end flower garden grow grow grow hand heart knew knew know know looked mother mrs oh old passed playing plucked put ran two two two wendy wendy years
1 beginning 1 children 1 cried 1 darling 1 day 1 delightful 1 end 1 flower 1 garden 3 grow 1 hand 1 heart 2 knew 2 know 1 looked 1 mother 1 mrs 1 oh 1 old 1 passed 1 playing 1 plucked 1 put 1 ran 1 remain 1 soon 1 subject 1 suppose 3 two 2 wendy 1 years
command join are commands for examining lines in sorted files
This is the working invocation
Also useful are sort and sort -rn

Function words

- Function words account for up to 60% of the content of a text
  - determiners: articles (the, a), possessive pronouns (their, your), quantifiers (much), demonstratives (that, those), and numbers
  - conjunctions: and, but, for, yet, neither, or, so, when, although, however, as, because, before
  - prepositions: in, of, between, on, with, by, at, without, through, over, across, around, into, within
  - pronouns: replacements for nouns: she, they, he, it, him, her, you, me, anybody, somebody, someone, anyone
  - auxiliary verbs: be, is, am, are, have, has, do, does, did, get, got, was, were
  - modals: express condition or possibility: may, might, can, could, will, would, shall, should
  - qualifiers: very, really, quite, somewhat, rather, too, pretty (much)
  - question words: how, where, what, when, why, who

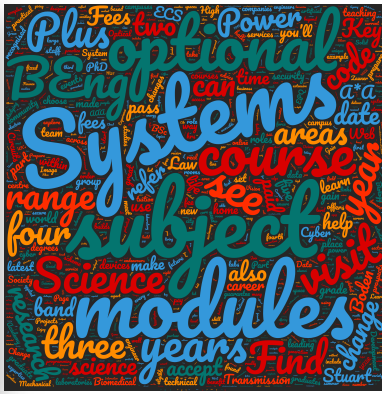
Use of Vocabularies

- Vocabularies are powerful things.
  - Counting words gives analytical insight
  - because repetition signals intention

In writing a speech, repetition is the key to leaving an impression. **Hammer home key words, phrases, and themes. Always be looking for places to tie back and reinforce earlier points**. And repeat critical points as if they were a musical refrain.

10 Keys To Writing A Speech - Forbes

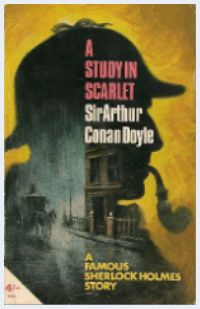
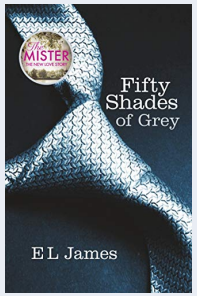
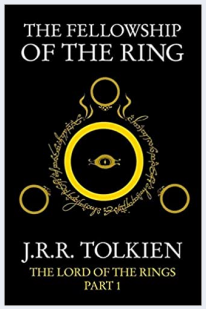
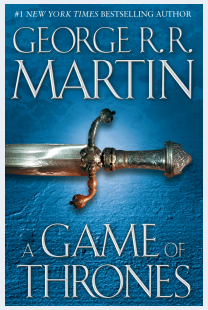
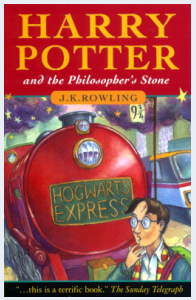
F www.forbes.com/sites/jeffschmitt/2013/07/16/10-keys-to-writing-a-speech/



ECS Prospectus

  - Wordclouds provide visual representation of statistical summary
    - great place to START research
    - see wordclouds.com or voyant-tools.org

Break Question: Can You Match the Right Book Covers to High Frequency Words?

	said cat people street know couldn name saying man eyes thought owls looked just cloak boy think seen little last day wall son sister say number corner went two turned suddenly sir right reason night mind looking four found fell drills called			said man little know asked knowledge companion fellow test rooms room answered round found blood appeared think stains practical mind long life hand good day work thought remarked question morning made looking idea hands far eyes expression doubt don't come came
	eyes don't says people know door smile he's white just glass questions make look interview hand work voice say office head hair good gaze think blonde young two things smiles recorder looks long gray floor control			cold said black sword man dead saw snow long hard fire eyes wind men made cloak came old lordling longsword light hand half away wood turned tree seen sable ridge moonlight looked heard hands felt asked
	said mother eyes looked just family children thought make good two take bit think society sister paper ball years ton say quite men man husband head father			said old say away came know come don't went ring door long thought course things guests birthday years good young quite presents people party mean large really looking called wonder right present place make made left keep joke folk eyes

These word lists come from the first chapter of each novel; each word appears 6 or more times.

Text Analysis First Step: Tokenization

- How to turn a stream of “characters” into a sequence of “words”
  - A **token** is a lexical construct (symbol) that can be assigned grammatical and semantic roles
  - How the token is constructed determines how the grammar and semantics can be understood

- Not about characters
  - Even defining a “character” is complex – 1,2,3,4 bytes?
  - Emojis are out 😢😂😭
  - Unicode, UTF-8, codepoints are mainly too complex £ “ ” – é
  - Stick to ASCII (character values 0..127) is safest
  - Processing ligatures is difficult ﬃ
  - Online book mixes up ”curly quotes” with ASCII quotes in its code examples
  - YOU CANNOT CUT AND PASTE PYTHON WITHOUT A SYNTAX ERROR
    - the word processor changed var= "string" to var=“string”

- Naïve solution: break on spaces & punctuation
  - too simple for general case
  - useful piece of information for parsers, help indicate sentence boundaries.

Tokenization Issues

- Internal punctuation
  - Abbreviations `m.p.h. Ph.D. AT&T cap'n`
  - prices `$45.55`
  - dates `01/02/2021`
  - times `12:34`
  - URLs `http://www.ecs.soton.ac.uk/`
  - hashtags `#nlproc`
  - social media tags `@lescarr`
  - email addresses `someone@cs.colorado.edu`
  - HTML entities `&`
  - Citations `Carr (1990)`
- Number expressions
  - `1,555,500.50` vs `1.555.500,50`
  - `3 and a half billion dollars`
  - `1.5 lbs`
- Clitic contractions
  - `we're I'm they'll, you've doesn't`
- Multiword expressions
  - `New York`
  - `rock 'n' roll`

Pattern Tokenization

```
>>> text = 'That U.S.A. poster-print costs $12.40...'
>>> pattern = r'''(?x)    # set flag to allow verbose regexps
...     ([A-Z]\.)+        # abbreviations, e.g. U.S.A.
...   | \w+(-\w+)*        # words with optional internal hyphens
...   | \$?\d+(\.\d+)?%?  # currency and percentages, e.g. $12.40, 82%
...   | \.\.\.            # ellipsis
...   | [][.,;"'?():-_`]  # these are separate tokens; includes ], [
... '''
>>> nltk.regexp_tokenize(text, pattern)
['That', 'U.S.A.', 'poster-print', 'costs', '$12.40', '...']
```

Figure 2.12: Speech and Language Processing Jurafsky & Martin 2020

- Use regular expression or other pattern matching style to define tokenization rules
  - Rules will be corpus specific and probably iteratively overfitted to painfully achieve sufficient accuracy for the task!
- Next lecture will be devoted to regular expressions

Tokenization

- Another way to tokenize text is to learn common patterns from the corpus itself
  - or from a similar “training corpus”
- By splitting words into *subword* units
  - Morphemes
  - Significant punctuation
- In a collection of research papers
  - `et al.` is a common bibliographic abbreviation
  - `O'Hara` is a common author surname
  - e.g. and i.e. are common formal abbreviations

Splitting Words - Lemmatization

- Determining words with different superficial forms but same root
  - cat -> cats, ate -> eaten, child -> children
- Words are composed of subword units called *morphemes*
- word parts with recognized meanings
  stems + affixes (prefixes or suffixes)
    - anti- in- ex- pre- post- un-
    - -able -al -ed -er -est -ing -ly -ness -tion
  - unfairness => un + fair + ness
- Can we find these common morphemes automatically?

Stemming

- Stemming is a (historical) simplistic lemmatization process that removes suffixes by applying a sequence of rewrites.
- The Porter stemmer applied to the following paragraph:
  - `This was not the map we found in Billy Bones's chest, but an accurate copy, complete in all things-names and heights and soundings-with the single exception of the red crosses and the written notes.`
- produces the following stemmed output:
  - `Thi wa not the map we found in Billi Bone s chest but an accur copi complet in all thing name and height and sound with the singl except of the red cross and the written note`
- The algorithm is a series of rewrite rules, in which the output of each pass is fed as input to the next pass:
  - ATIONAL → ATE (e.g., relational → relate)
  - ING → if stem contains vowel (e.g., motoring → motor)
  - SSES → SS (e.g., grasses → grass)

- Very language-specific, very crude

BPE Byte Pair Encoding

- Byte-Pair Encoding Algorithm start with a vocabulary that is just the set of all individual characters [a, b, … z, <endword>] and tries to learn *k* new tokens

- Repeat *k times* examining the corpus
  - select the two symbols that are most frequently adjacent (A, B),
  - add a new merged symbol AB to the vocabulary,
  - replace every adjacent A B in the corpus with the new AB

    see figure 2.13 from *Speech and Language Processing Jurafsky & Martin 2020*

- BPE is run with many thousands of merges on a very large input corpus.
  - Most words will be represented as full symbols
  - only the very rare words (and unknown words) will have to be represented by their parts.

BPE worked example

START: vocab = a-z + word-end

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w
2 l o w e s t _	
6 n e w e r _	
3 w i d e r _	
2 n e w _	

Most common pair: r _

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w, er
2 l o w e s t _	
6 n e w er _	
3 w i d er _	
2 n e w _	

Most common pair: e r_

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w, er, er_
2 l o w e s t _	
6 n e w er_	
3 w i d er_	
2 n e w _	

Most common pair: n e

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne
2 l o w e s t _	
6 ne w er_	
3 w i d er_	
2 ne w _	

_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new

_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo

_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low

_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_

_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_

- Keep going until you have *enough symbols*

BPE, Wordpiece & SentencePiece

- BPE defines a generalised process
  - iteratively extending a set of tokens by looking for the “best performing” new token

- BPE is the simplest of a set of algorithms that use different conceptions of “token” and “best performing” to improve performance on very large, highly diverse datasets

- **WordPiece** (see section 11.7.1 of *Jurafsky & Martin 2020)* is based on an n-gram language model using multiple-adjacent words as single tokens (see later lectures for details)

- **SentencePiece** extends these as a simple and language independent text tokenizer (mainly for neural network-based text generation systems).
  - Crucially, it is an algorithm does not depend on any language-specific processing, handling European and Asian languages equivalently.

Word Similarity

- How similar are two words?
  - accommodate *vs* accomodate — One letter deleted
  - achieve *vs* acheive — Two letters swapped
    - **Common spelling errors**
  - Ich bin ein Berliner *vs* Ich bin Berliner — One word deleted
    - **Misquoted JFK speech after the Berlin wall erected in 1963**
  - Excuse me while I kiss this guy *vs* Excuse me while I kiss the sky — Three letters deleted + two inserted
    - **Commonly misheard lyrics by Jimi Hendrix (1967)**

Levenshtein distance

- Shortest sequence of edits to transform one string into another
- Metric of the similarity of two strings

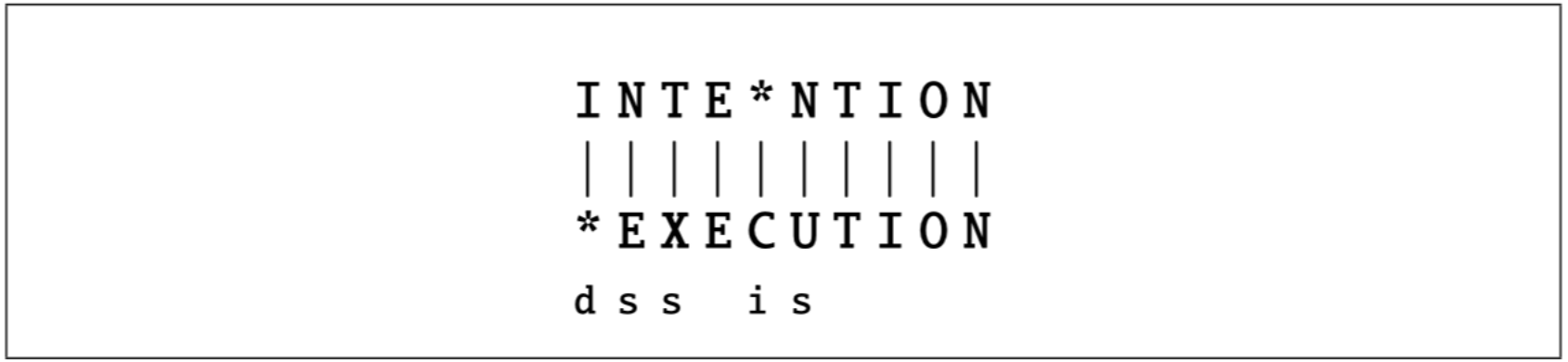


Figure 2.14: Speech and Language Processing Jurafsky & Martin 2020

- Alignment diagram demonstrates that 5 operations are required
  - 3 substitutions, a delete and an insert
  - Each have a cost of 1 (operation)
    - Alternatively, substitution = delete + insert at cost of 2

Minimum Edit Distance

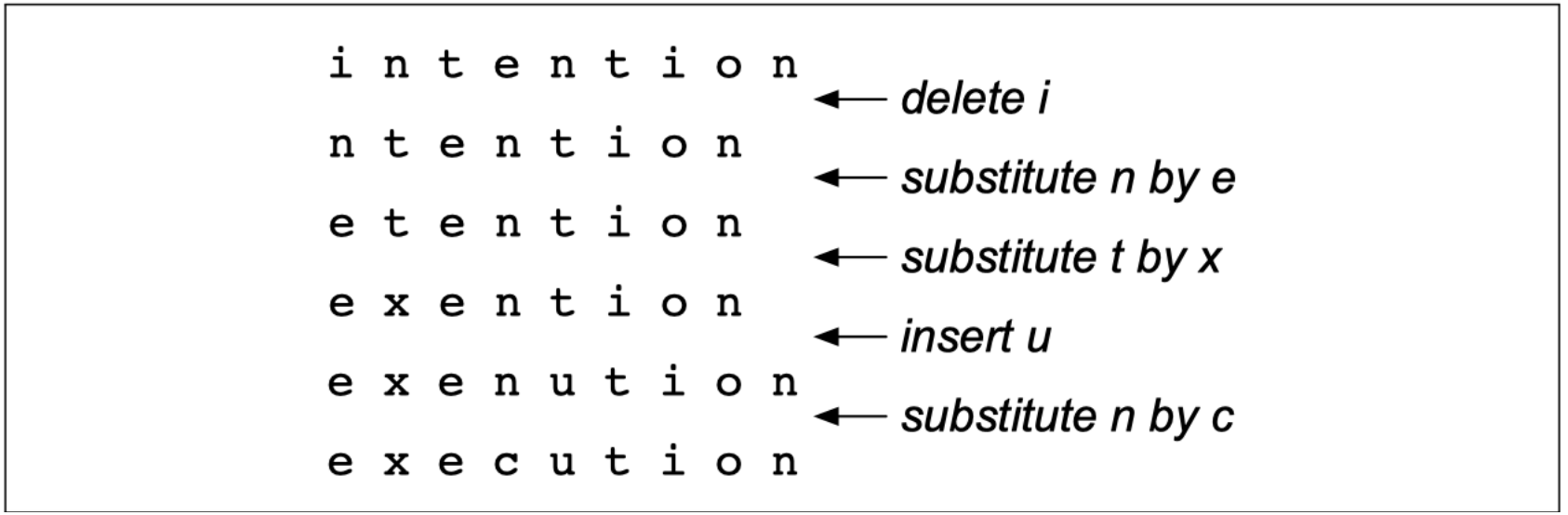


Figure 2.16: Speech and Language Processing Jurafsky & Martin 2020

- Applications
  - similarities in diffs generation for software updates (versions)
  - plagiarism analysis
  - alignments in parallel corpora (e.g. translations)

Calculating the Levenshtein metric

- Can be calculated recursively

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}(\operatorname{tail}(a), b) \\ \operatorname{lev}(a, \operatorname{tail}(b)) \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

insert

delete

replace

where the *tail* of some string x is a string of all but the first character of x, and *x[n]* is the nth character of the string x, starting with character 0.

From Wikipedia entry "Levenshtein distance"
See also pp22-26 of Speech and Language Processing Jurafsky & Martin 2020

End of Lecture Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

Please spend **1 minute** using Panopto quiz to wrie down two or three questions that you would like to have answered at the next interactive session.

Do it **right now** while its fresh.

Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.