Text Preprocessing in Natural Language Processing

Transforming Raw Text into Actionable Data

Announcements

- Lab session: Thursday 4 - 6 pm **Computer Lab-3** [Next week onwards]
- Tuesday 2 - 2:50 pm (CS-1) will be used for extra class or evaluation activities.

Lab evaluation:

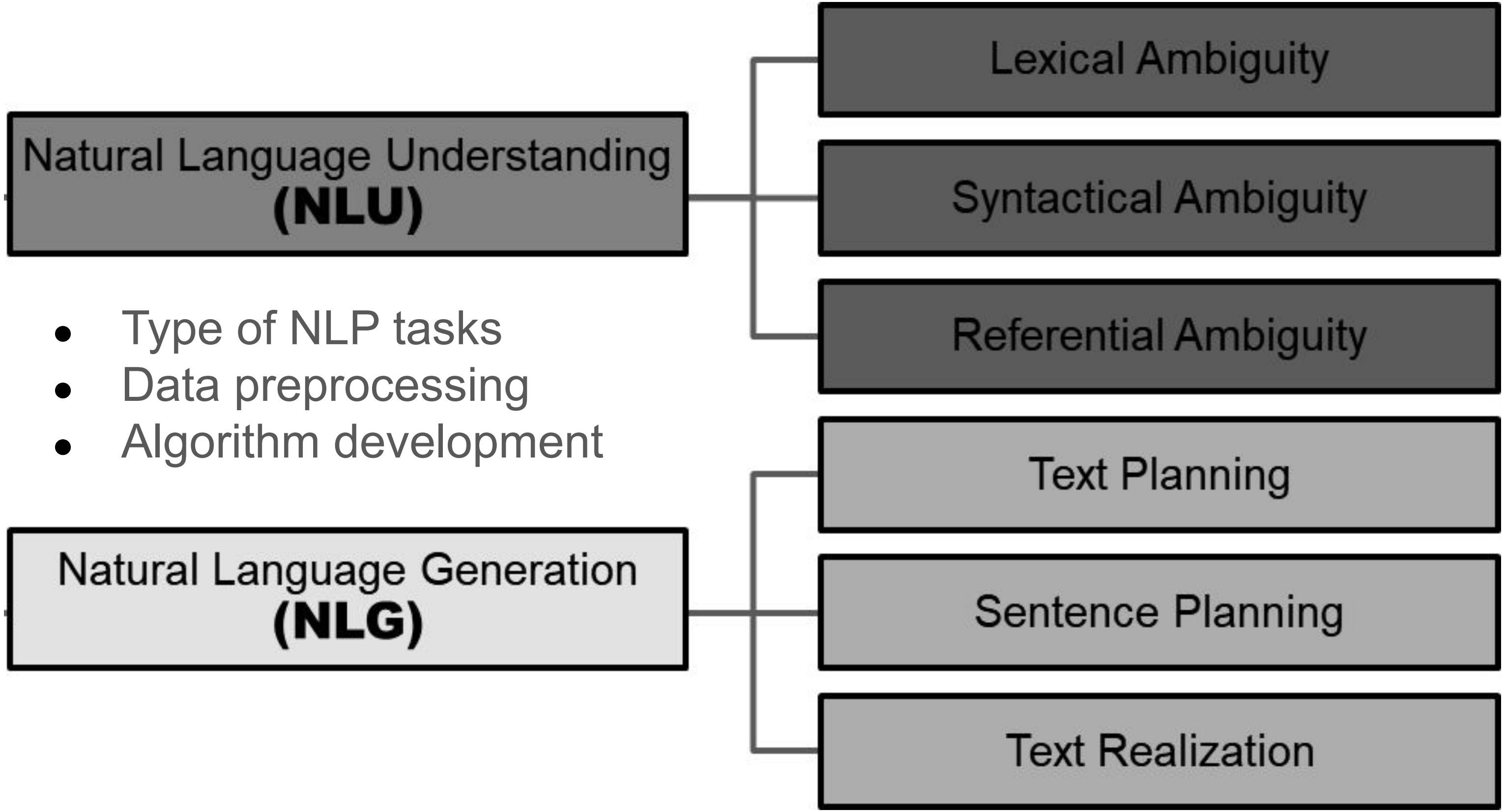
- 2 hours of practical. [Topics will be shared during the class lectures]
- 1 hour evaluation [Will announce the date later]
- Assignment 1 evaluation will be done on **Tuesday (05/08/2025)**
- Enroll to Google Classroom

Exam

- Assessment is by 100% in-person examination
- Weightage distribution:
  - Lab evaluation: 15%
  - Quizzes: 10%
  - Mid-sem written exam: 20%
  - End-sem written exam: 30%
  - Project based assessment: 25%

Quick recap of previous lecture

Components of NLP



Natural Language Understanding **(NLU)**

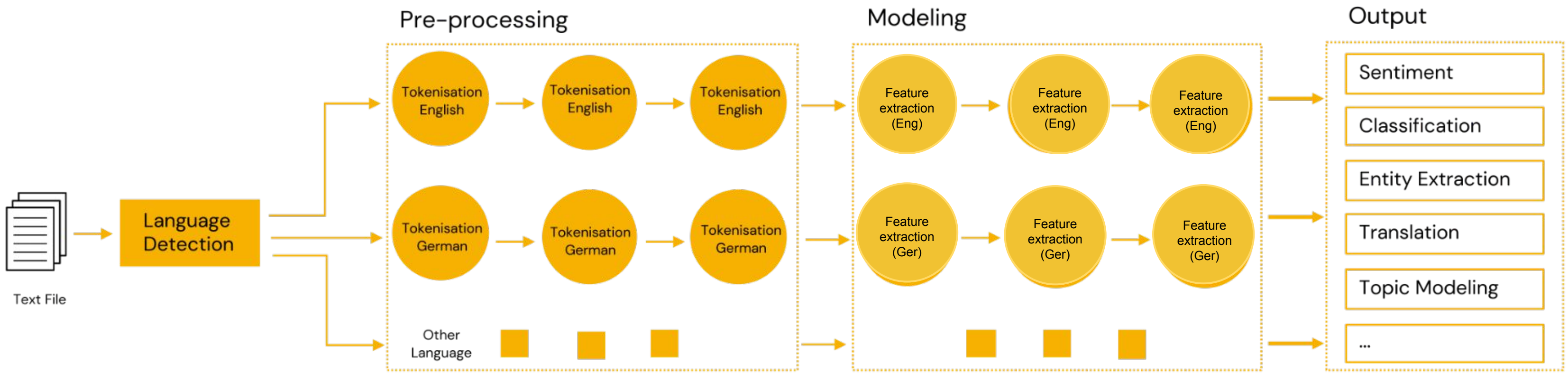
- Lexical Ambiguity
- Syntactical Ambiguity
- Referential Ambiguity

- Type of NLP tasks
- Data preprocessing
- Algorithm development

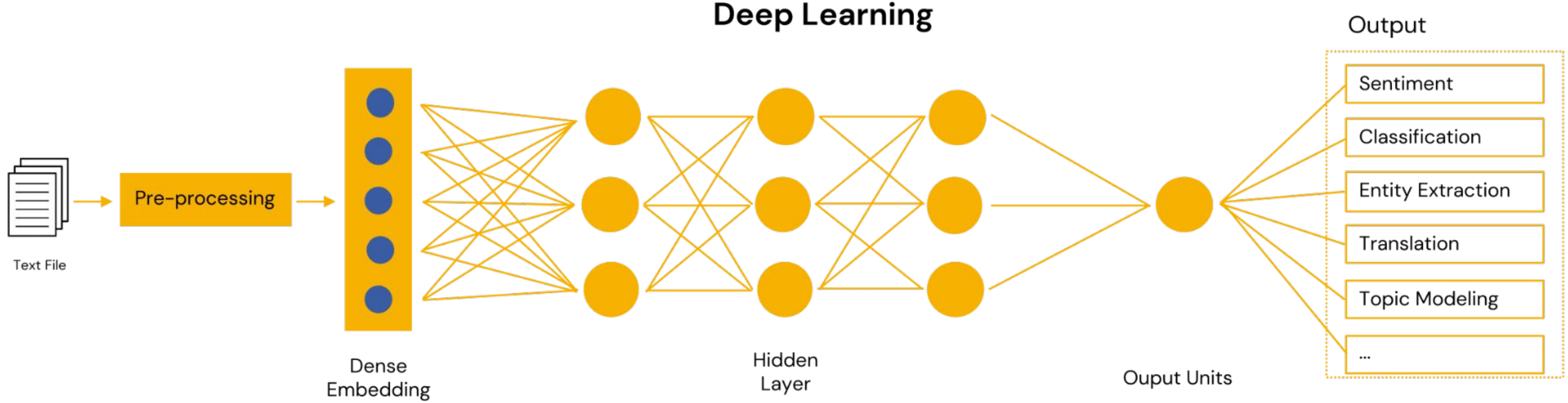
Natural Language Generation **(NLG)**

- Text Planning
- Sentence Planning
- Text Realization

Classical NLP



Deep Learning



Outline

- Words
- What is text preprocessing?
- Why is it important?
- Preprocessing techniques

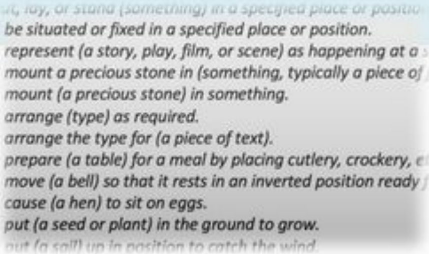
Computers

- Unambiguous
- Instructions
- Expected behaviour
- Modular
- Clear
- Precise

```
ws.on("message", m => {
  let a = m.split(" ")
  switch(a[0]){
    case "connect":
      if(a[1]){
        if(clients.has(a[1])){
          ws.send("connected");
          ws.id = a[1];
        }else{
          ws.id = a[1]
          clients.set(a[1], {client: {position:{x: 0, y: 0},
          ws.send("connected")
        }
      }else{
        let id = Math.random().toString().slice(2, 10)
```

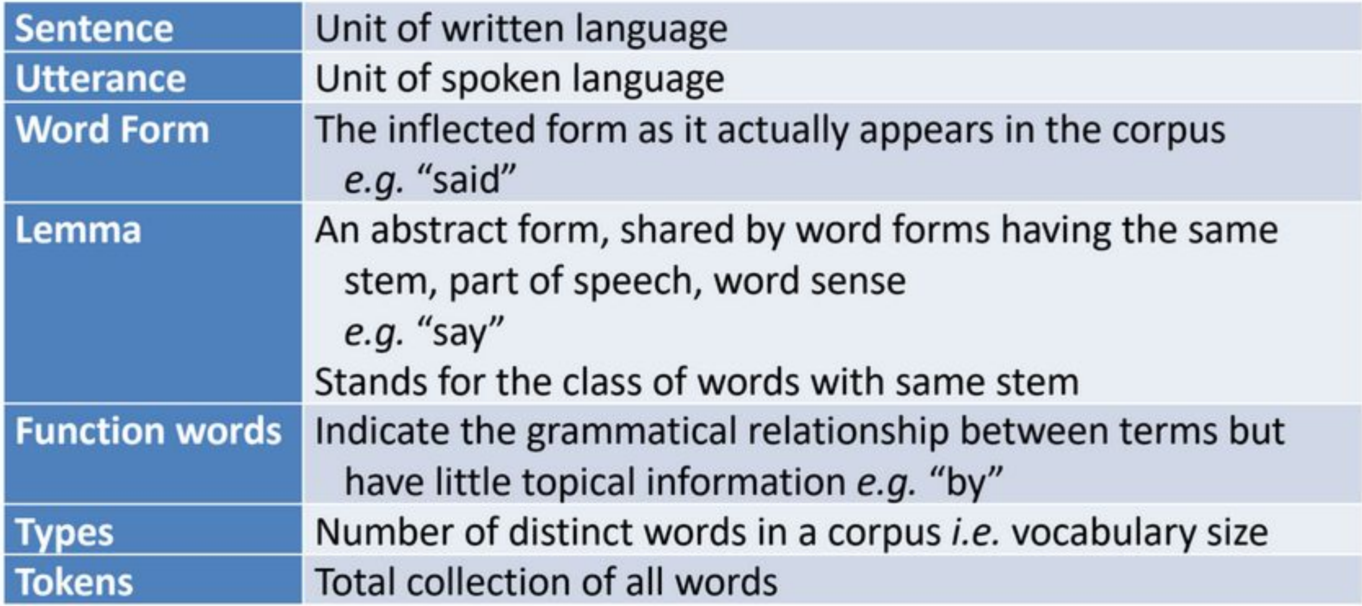
Humans

- Depends.. Punctuation, intonation, emphasis
- Vague, abstract
- Multiple senses of a single word
- Context
- Listener interpretation

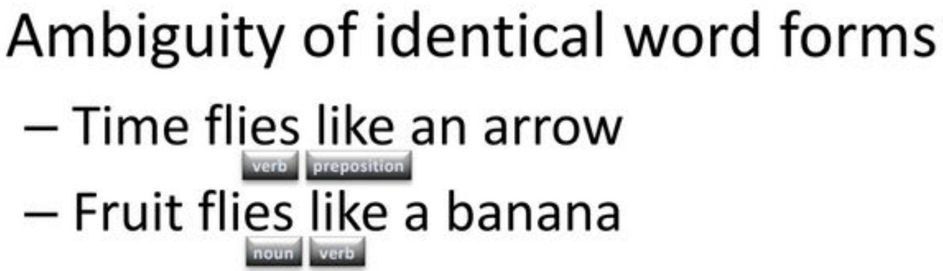


put, lay, or stand (something) in a specified place or positio
be situated or fixed in a specified place or position.
represent (a story, play, film, or scene) as happening at a
mount a precious stone in (something, typically a piece of
mount (a precious stone) in something.
arrange (type) as required.
arrange the type for (a piece of text).
prepare (a table) for a meal by placing cutlery, crockery, e
move (a bell) so that it rests in an inverted position ready
cause (a hen) to sit on eggs.
put (a seed or plant) in the ground to grow.
put (a sail) up in position to catch the wind.

Useful terminology



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



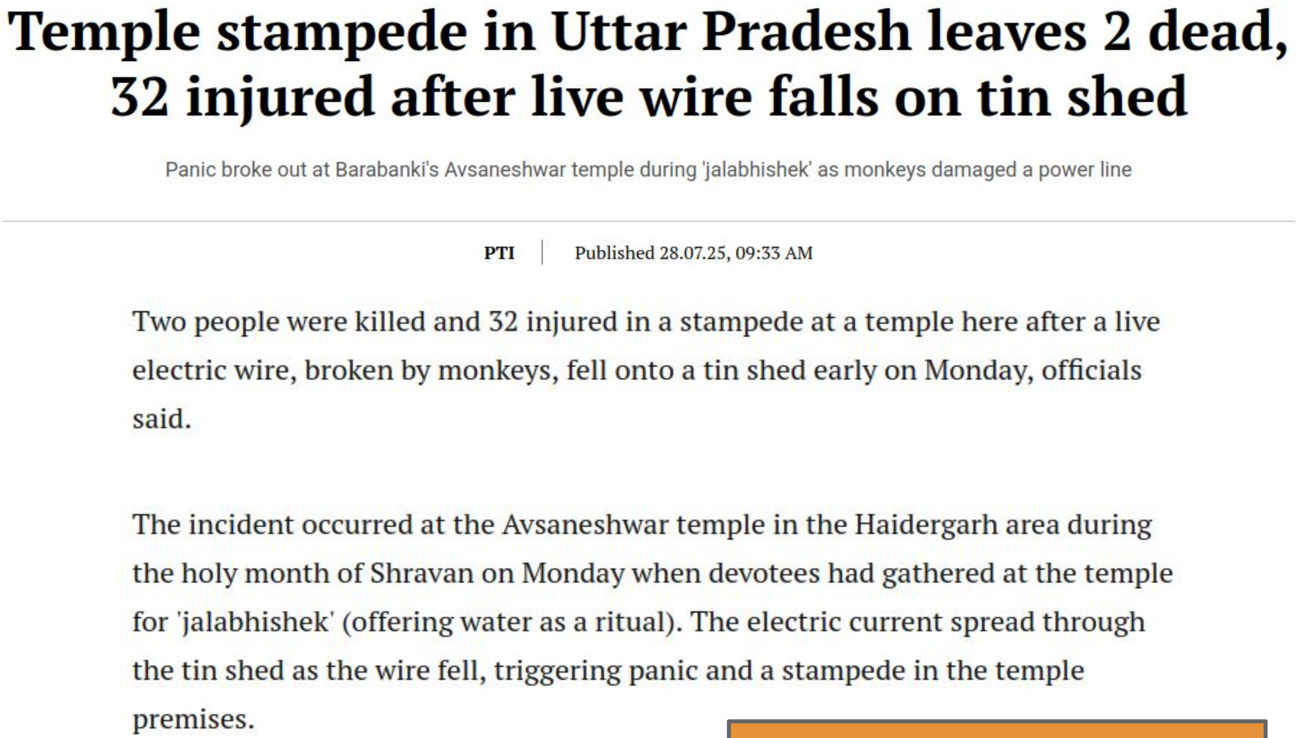
Ambiguity of identical word forms

- Time flies like an arrow
  verb preposition
- Fruit flies like a banana
  noun verb

[1] Jawaan is all about **heavy** action. Loved it!!!

[2] My new phone's so **heavy**, it's practically a hammer!

It’s only words:



Temple stampede in Uttar Pradesh leaves 2 dead, 32 injured after live wire falls on tin shed

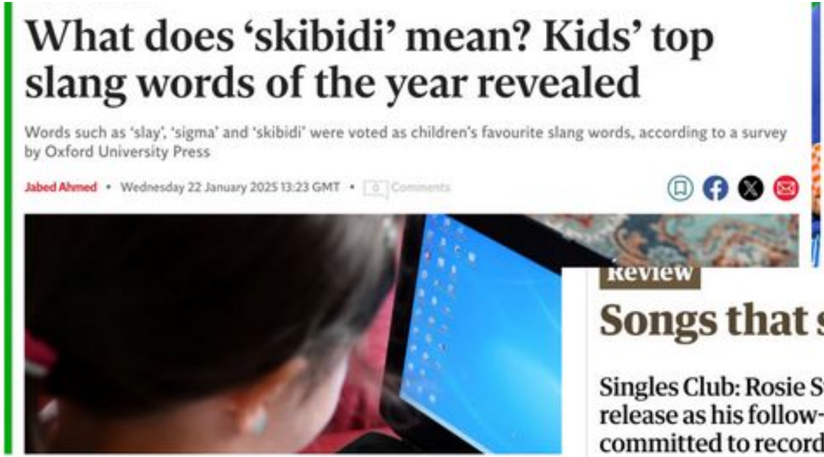
Panic broke out at Barabanki's Avsaneshwar temple during 'jalabhishek' as monkeys damaged a power line

PTI | Published 28.07.25, 09:33 AM

Two people were killed and 32 injured in a stampede at a temple here after a live electric wire, broken by monkeys, fell onto a tin shed early on Monday, officials said.

The incident occurred at the Avsaneshwar temple in the Haidergarh area during the holy month of Shravan on Monday when devotees had gathered at the temple for 'jalabhishek' (offering water as a ritual). The electric current spread through the tin shed as the wire fell, triggering panic and a stampede in the temple premises.

www.telegraphindia.com



What does ‘skibidi’ mean? Kids’ top slang words of the year revealed

Words such as ‘slay’, ‘sigma’ and ‘skibidi’ were voted as children’s favourite slang words, according to a survey by Oxford University Press

Jabed Ahmed • Wednesday 22 January 2025 13:23 GMT • 0 Comments

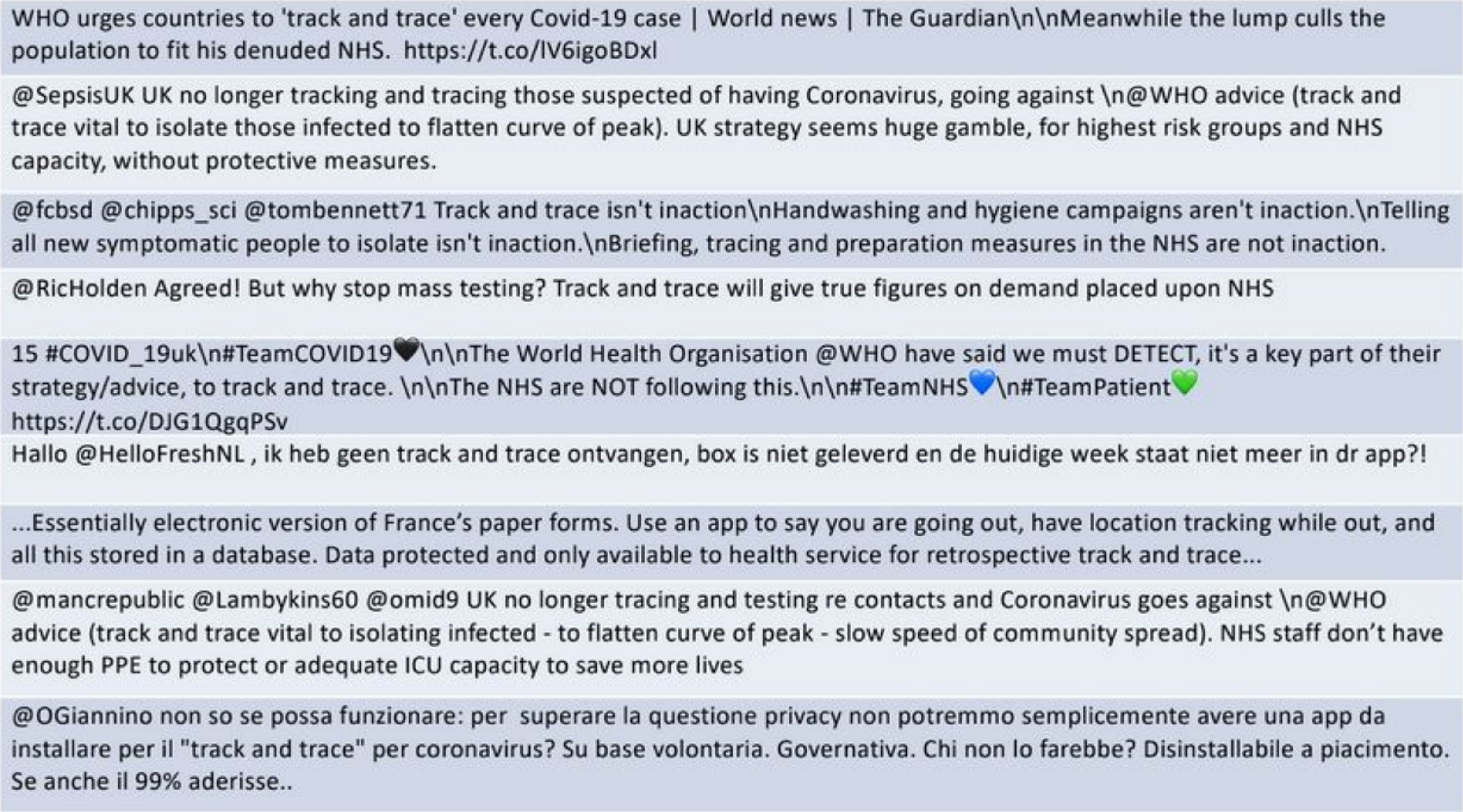
Controversial (Independent)

Review

Songs that say 'Up yours, Radio 1!'

Singles Club: Rosie Swash admires Sam Sparro's decision to release as his follow-up single one of the worst songs ever committed to record

Controversial (Guardian)

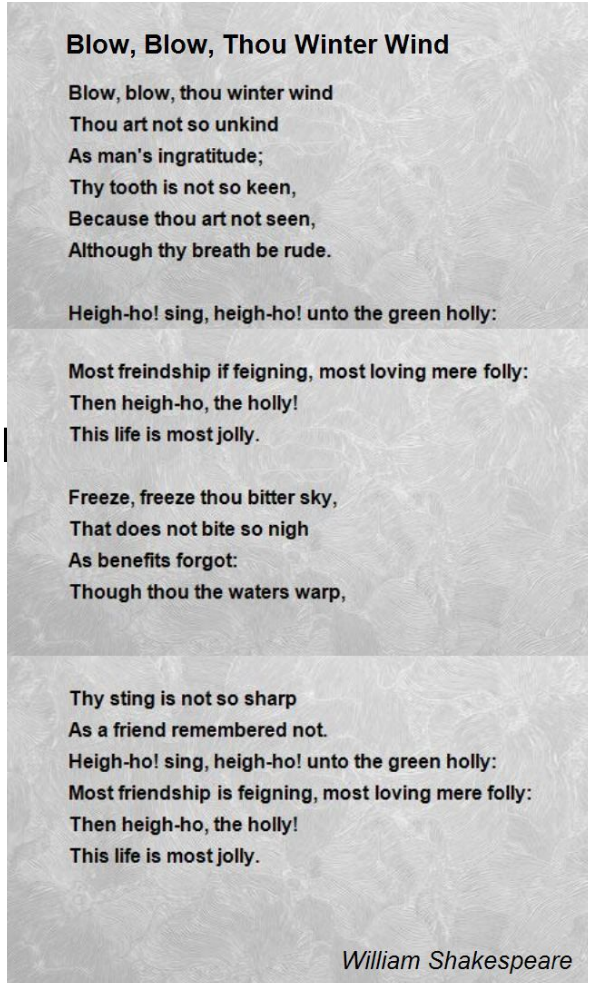


WHO urges countries to 'track and trace' every Covid-19 case | World news | The Guardian\n\nMeanwhile the lump culls the population to fit his denuded NHS. https://t.co/lV6igoBDxl
@SepsisUK UK no longer tracking and tracing those suspected of having Coronavirus, going against \n@WHO advice (track and 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..

Language use, context, and meaning

- Formal prose – news, essays, official reports
- Casual news – satire, humour, entertainment
- Speech
- Poetry
- Music
- Social media

“I do uh main- mainly business data



Blow, Blow, Thou Winter Wind

Blow, blow, thou winter wind
Thou art not so unkind
As man's ingratitude;
Thy tooth is not so keen,
Because thou art not seen,
Although thy breath be rude.

Heigh-ho! sing, heigh-ho! unto the green holly:

Most freindship if feigning, most loving mere folly:
Then heigh-ho, the holly!
This life is most jolly.

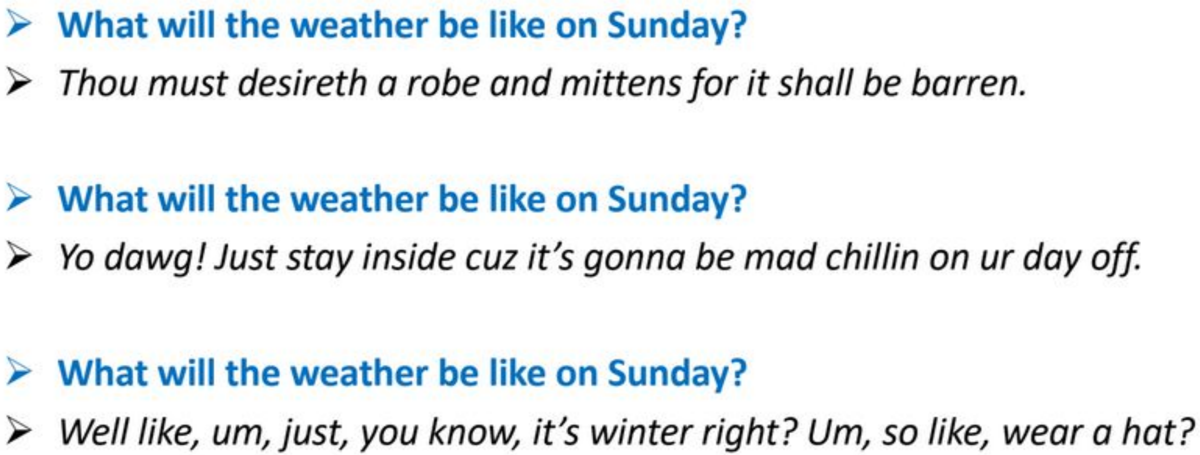
Freeze, freeze thou bitter sky,
That does not bite so nigh
As benefits forgot:
Though thou the waters warp,

Thy sting is not so sharp
As a friend remembered not.
Heigh-ho! sing, heigh-ho! unto the green holly:
Most friendship is feigning, most loving mere folly:
Then heigh-ho, the holly!
This life is most jolly.

William Shakespeare

Why does the writing style matter?

Imagine ChatGPT was trained on a certain type of text styles…

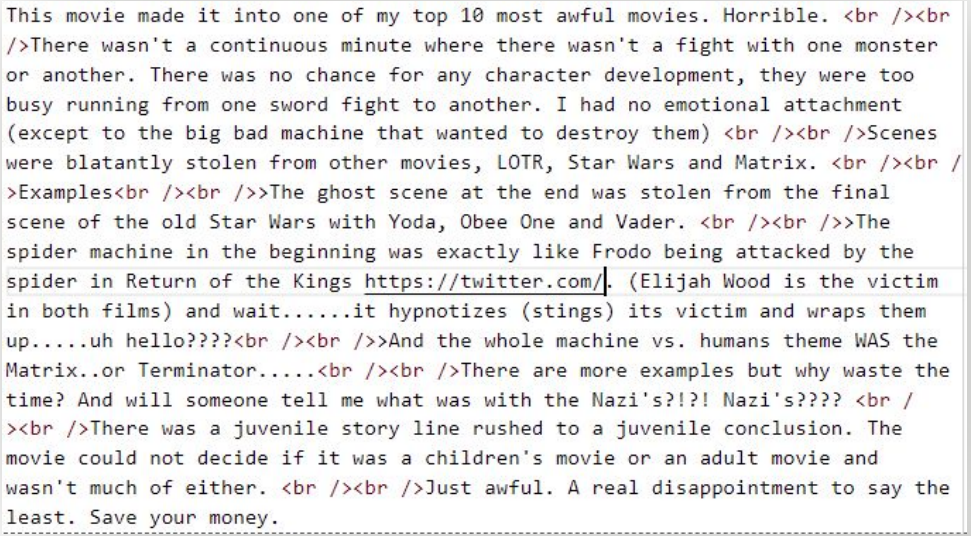


- **What will the weather be like on Sunday?**
- *Thou must desireth a robe and mittens for it shall be barren.*

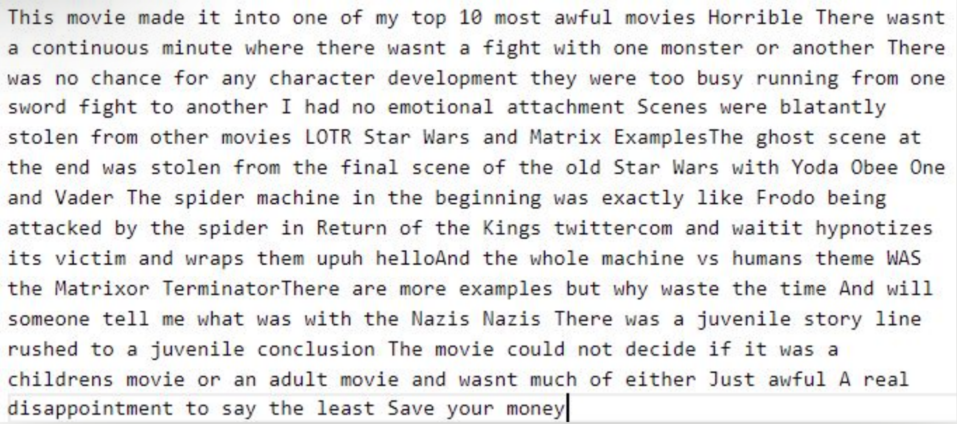
- **What will the weather be like on Sunday?**
- *Yo dawg! Just stay inside cuz it's gonna be mad chillin on ur day off.*

- **What will the weather be like on Sunday?**
- *Well like, um, just, you know, it's winter right? Um, so like, wear a hat?*

Cleaning web crawled data



```
This movie made it into one of my top 10 most awful movies. Horrible. <br /><br
/>There wasn't a continuous minute where there wasn't a fight with one monster
or another. There was no chance for any character development, they were too
busy running from one sword fight to another. I had no emotional attachment
(except to the big bad machine that wanted to destroy them) <br /><br />Scenes
were blatantly stolen from other movies, LOTR, Star Wars and Matrix. <br /><br /
>Examples<br /><br />>The ghost scene at the end was stolen from the final
scene of the old Star Wars with Yoda, Obee One and Vader. <br /><br />>The
spider machine in the beginning was exactly like Frodo being attacked by the
spider in Return of the Kings https://twitter.com/. (Elijah Wood is the victim
in both films) and wait......it hypnotizes (stings) its victim and wraps them
up.....uh hello????<br /><br />>And the whole machine vs. humans theme WAS the
Matrix..or Terminator.....<br /><br />There are more examples but why waste the
time? And will someone tell me what was with the Nazi's?!?! Nazi's???? <br /
><br />There was a juvenile story line rushed to a juvenile conclusion. The
movie could not decide if it was a children's movie or an adult movie and
wasn't much of either. <br /><br />Just awful. A real disappointment to say the
least. Save your money.
```



```
This movie made it into one of my top 10 most awful movies Horrible There wasnt
a continuous minute where there wasnt a fight with one monster or another There
was no chance for any character development they were too busy running from one
sword fight to another I had no emotional attachment Scenes were blatantly
stolen from other movies LOTR Star Wars and Matrix ExamplesThe ghost scene at
the end was stolen from the final scene of the old Star Wars with Yoda Obee One
and Vader The spider machine in the beginning was exactly like Frodo being
attacked by the spider in Return of the Kings twittercom and waitit hypnotizes
its victim and wraps them upuh helloAnd the whole machine vs humans theme WAS
the Matrixor TerminatorThere are more examples but why waste the time And will
someone tell me what was with the Nazis Nazis There was a juvenile story line
rushed to a juvenile conclusion The movie could not decide if it was a
childrens movie or an adult movie and wasnt much of either Just awful A real
disappointment to say the least Save your money
```

What is text preprocessing?

- **Definition:** Cleaning and structuring raw text for NLP tasks.
- **Goal:** Reduce noise, standardize formats, and extract meaningful features.

Why text preprocessing is important?

- **Noise Reduction:** Remove irrelevant characters (punctuation, HTML tags).
- **Consistency:** Lowercasing, standardizing formats (e.g., dates).
- **Efficiency:** Smaller vocabulary size = faster model training.
- **Accuracy:** Improves NLP task performance (e.g., sentiment analysis).

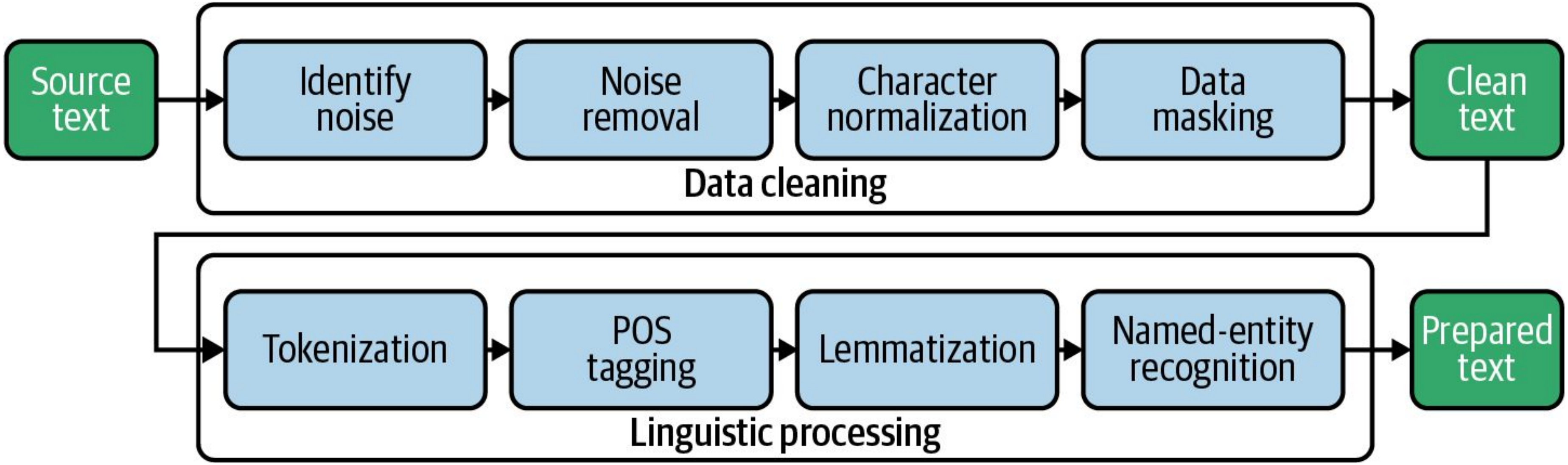
Challenges

- **Ambiguity:** "Apple" (fruit vs. company).
- **Language Differences:** Morphology in Arabic vs. English.
- **Resource Limits:** Lemmatization requires heavy dictionaries.

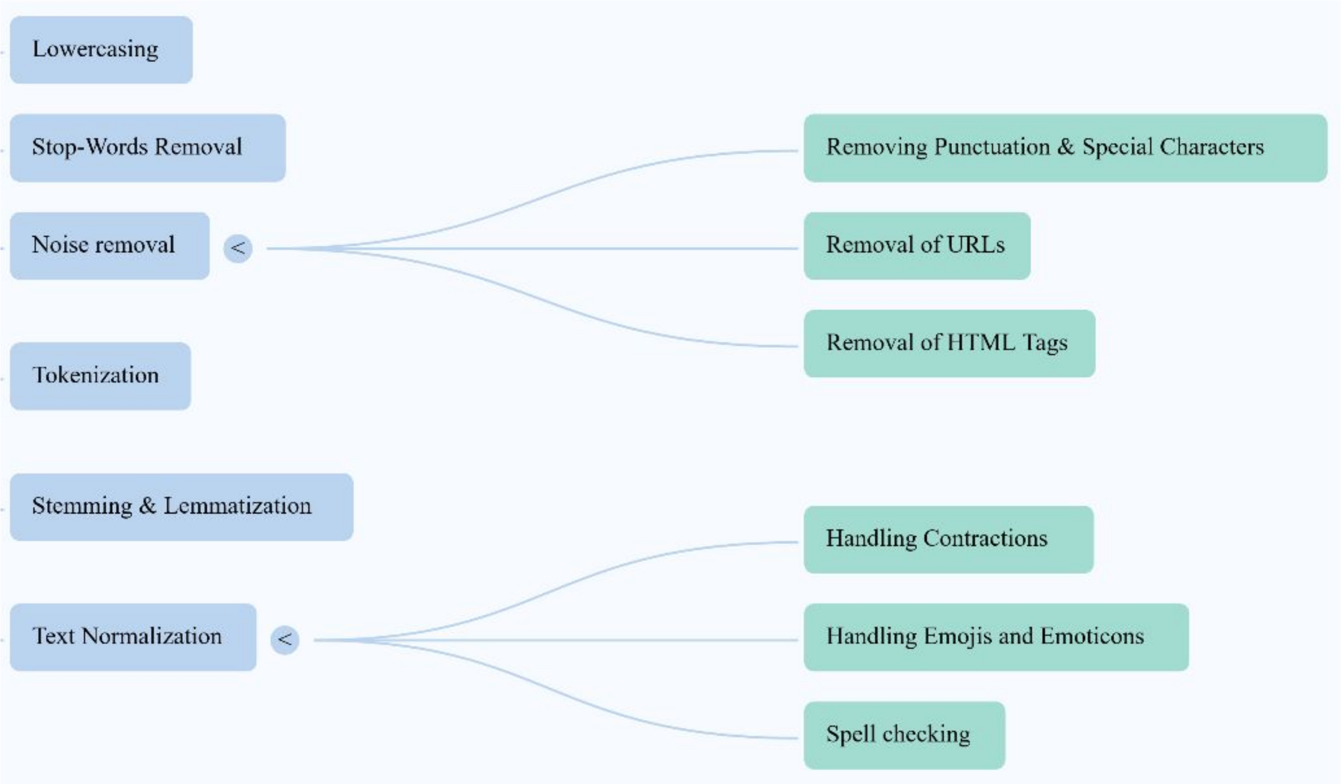
Domain-Specific Preprocessing

- **Social Media:** Emoji handling, slang normalization.
- **Scientific Texts:** Retain equations/symbols.
- **Medical Texts:** Protect sensitive terms (e.g., patient names).

Text preprocessing stages



Common preprocessing steps:



- Lowercasing
- Stop-Words Removal
- Noise removal <
  - Removing Punctuation & Special Characters
  - Removal of URLs
  - Removal of HTML Tags
- Tokenization
- Stemming & Lemmatization
- Text Normalization <
  - Handling Contractions
  - Handling Emojis and Emoticons
  - Spell checking

Text preprocessing techniques in NLP

- Regular expression
- Tokenization
- Lemmatization
- Stemming
- Part-of-speech (POS) tagging
- Name-Entity-Recognition (NER)

Toolkits:

- **NLTK:** Noise removal, Tokenization, stemming, POS/NER tagging.
- **spaCy:** Tokenization, stemming, POS/NER tagging, Dependency Parsing.
- **Hugging Face Tokenizers:** For transformer models.

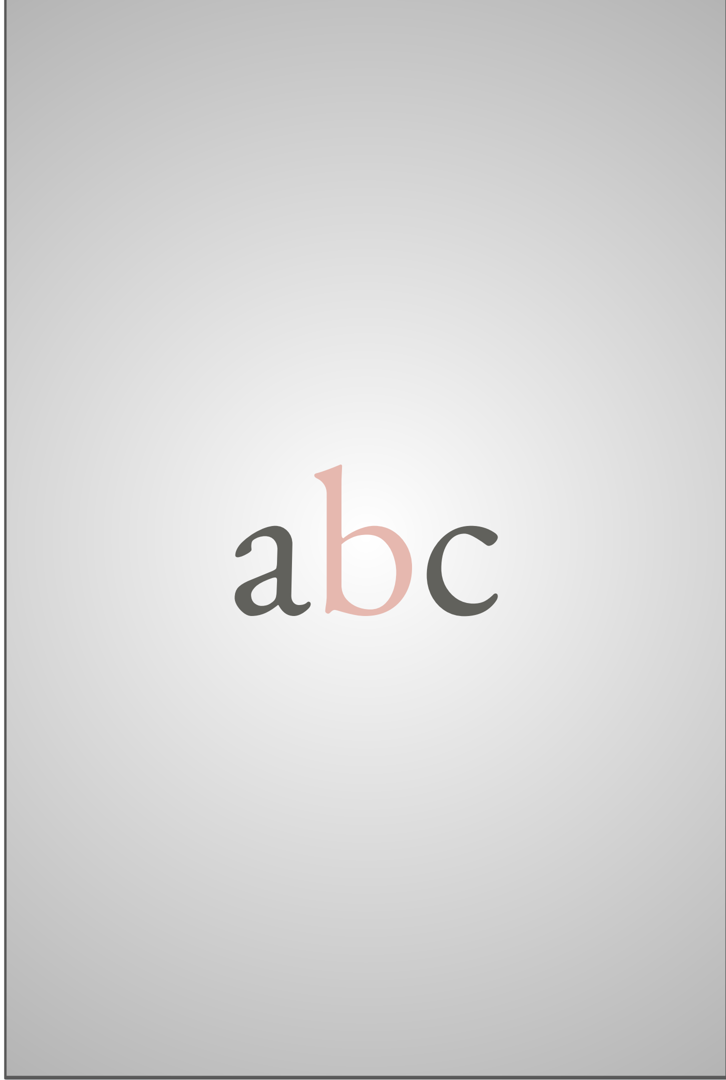
Lowercasing and Stop-Words Removal

Lowercasing

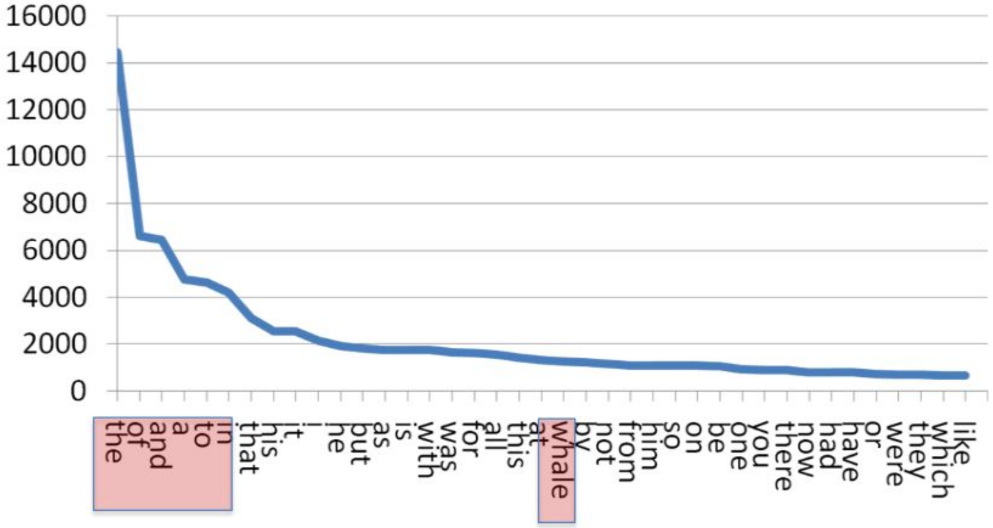
Convert all text to lowercase for uniformity

Stop-Words Removal

Eliminate common words that add little meaning

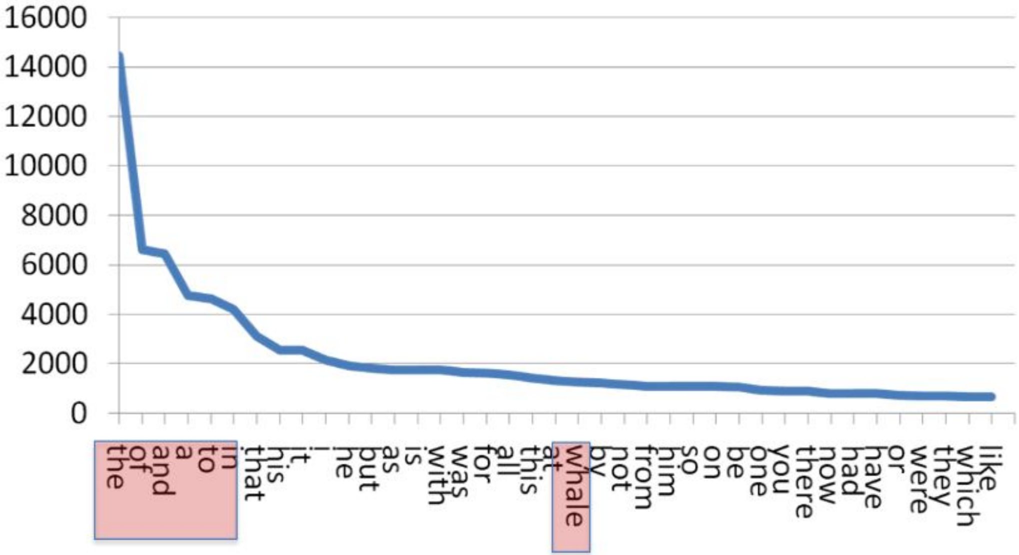


Word frequency follows a Zipfian distribution



the
and
to
of
in
we
for
that
with
father
your
is
you
this
most
his
it
be
our
nobility
business
toward
think
him
note
world
know
out
an
fault
dear
Wittenberg
sorrow
impart
earth
heart
death
common
have
lost
take
son
further
nature
much
going
cheer
let
retrograde
Norway
Hamlet
cousin
chiefest
all
are
here
heaven
dearest
us
no
less
bend
courtier
so
do
love
which
herein
than
eye
tis
as
desire
pray
back
beseech
woe
young
but
must
throne
remain
unprevailing
school
grief
comfort
by
brother
Fortinbras
throw
immediate
intent
who
hath
he
now
these
state
yet
bears

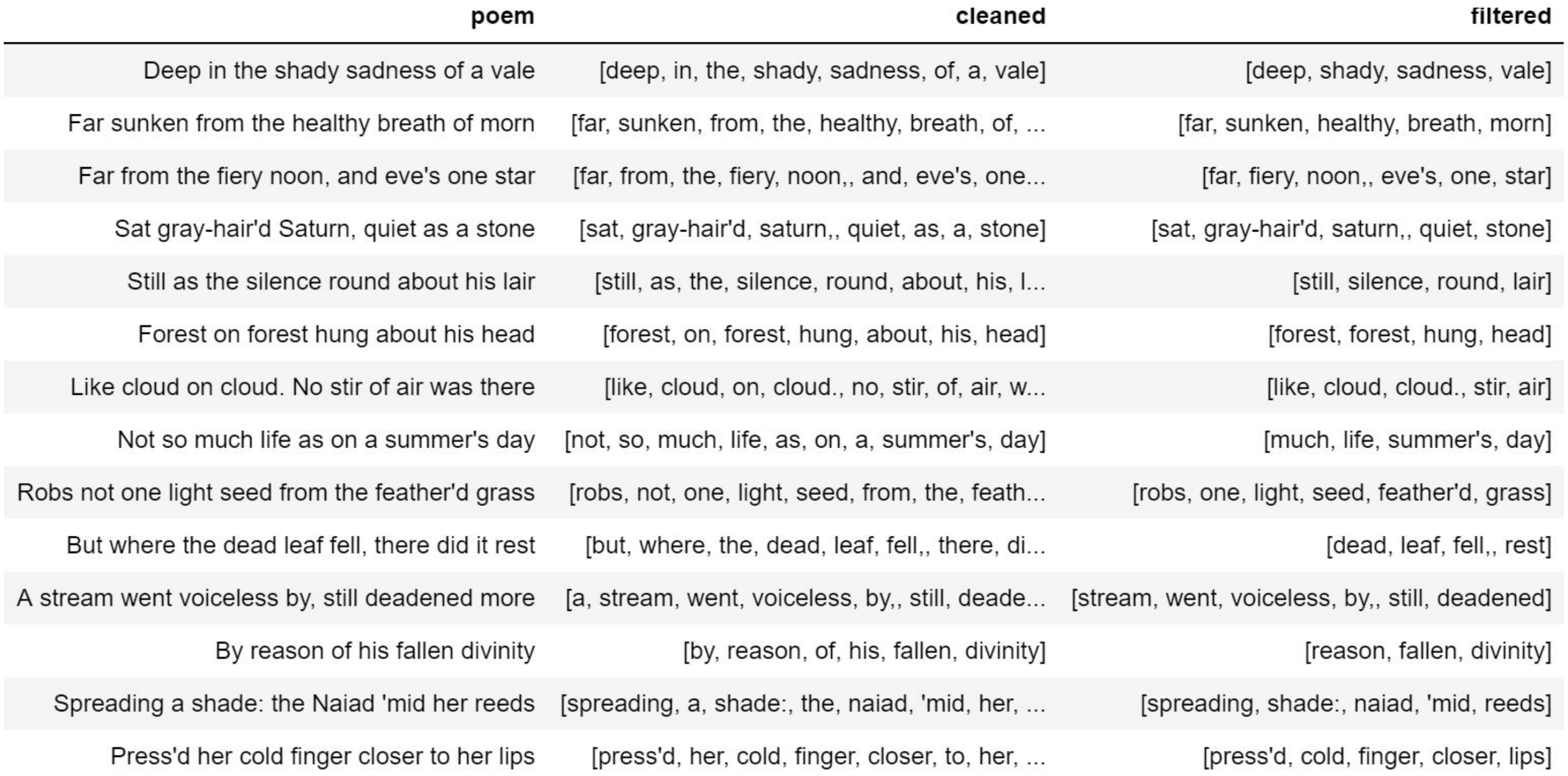
Word frequency follows a Zipfian distribution



Top frequent words are generally functional words:

- **Determiners**: a, an, the, you, them, that, those, etc.
- **Conjunctions**: and, but, for, …
- **Prepositions**: in, of, by, at, …
- **Auxiliary verbs**: be, is, am, was, were, do, does, …
- **Qualifiers**: very, really, too, pretty, rather, quite, …
- **Questions**: how, what, where, when, why, who

Stopword removal

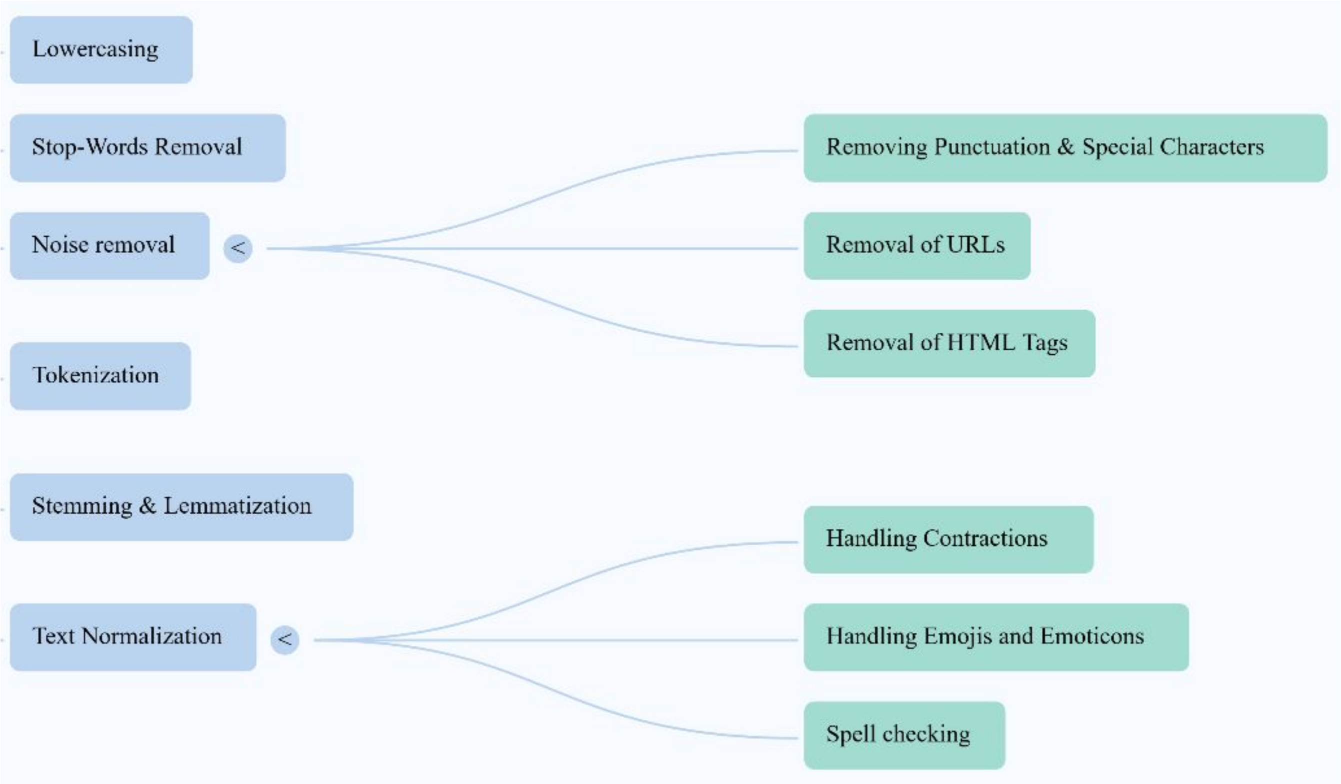


poem	cleaned	filtered
Deep in the shady sadness of a vale	[deep, in, the, shady, sadness, of, a, vale]	[deep, shady, sadness, vale]
Far sunken from the healthy breath of morn	[far, sunken, from, the, healthy, breath, of, ...	[far, sunken, healthy, breath, morn]
Far from the fiery noon, and eve's one star	[far, from, the, fiery, noon,, and, eve's, one...	[far, fiery, noon,, eve's, one, star]
Sat gray-hair'd Saturn, quiet as a stone	[sat, gray-hair'd, saturn,, quiet, as, a, stone]	[sat, gray-hair'd, saturn,, quiet, stone]
Still as the silence round about his lair	[still, as, the, silence, round, about, his, l...	[still, silence, round, lair]
Forest on forest hung about his head	[forest, on, forest, hung, about, his, head]	[forest, forest, hung, head]
Like cloud on cloud. No stir of air was there	[like, cloud, on, cloud., no, stir, of, air, w...	[like, cloud, cloud., stir, air]
Not so much life as on a summer's day	[not, so, much, life, as, on, a, summer's, day]	[much, life, summer's, day]
Robs not one light seed from the feather'd grass	[robs, not, one, light, seed, from, the, feath...	[robs, one, light, seed, feather'd, grass]
But where the dead leaf fell, there did it rest	[but, where, the, dead, leaf, fell,, there, di...	[dead, leaf, fell,, rest]
A stream went voiceless by, still deadened more	[a, stream, went, voiceless, by,, still, deade...	[stream, went, voiceless, by,, still, deadened]
By reason of his fallen divinity	[by, reason, of, his, fallen, divinity]	[reason, fallen, divinity]
Spreading a shade: the Naiad 'mid her reeds	[spreading, a, shade:, the, naiad, 'mid, her, ...	[spreading, shade:, naiad, 'mid, reeds]
Press'd her cold finger closer to her lips	[press'd, her, cold, finger, closer, to, her, ...	[press'd, cold, finger, closer, lips]

How to define stop words?

- **Statistical Methods**
  - **Frequency-Based Approach:** Calculate frequencies of all words and remove those appearing most often (top 20% or so). High-frequency words across documents often have little discriminative value.
  - **Inverse Document Frequency (IDF):** Words appearing in almost all documents (having very low IDF values) make good stopword candidates.
  - **TF-IDF Method:** Using the TF-IDF formula, words with scores of 0 or very close to 0 exist in all documents and likely carry little semantic value.

Common preprocessing steps:



Lowercasing

Stop-Words Removal

Noise removal

- Removing Punctuation & Special Characters
- Removal of URLs
- Removal of HTML Tags

Tokenization

Stemming & Lemmatization

Text Normalization

- Handling Contractions
- Handling Emojis and Emoticons
- Spell checking

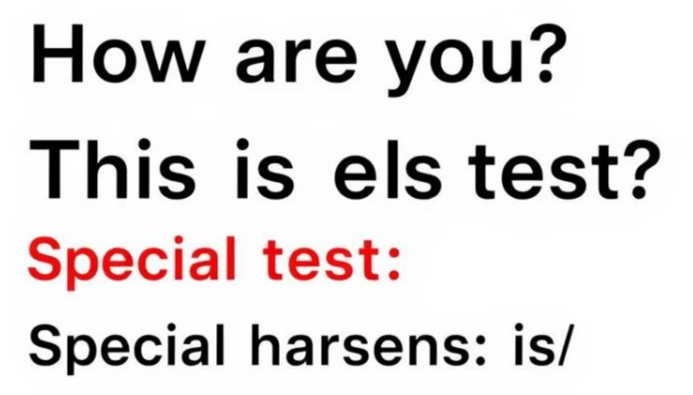
Noise Removal: Punctuation & Special Characters

1 Remove punctuation

Strip symbols like commas, periods, brackets

2 Remove special chars

Eliminate non-text symbols [including emojis and symbols]

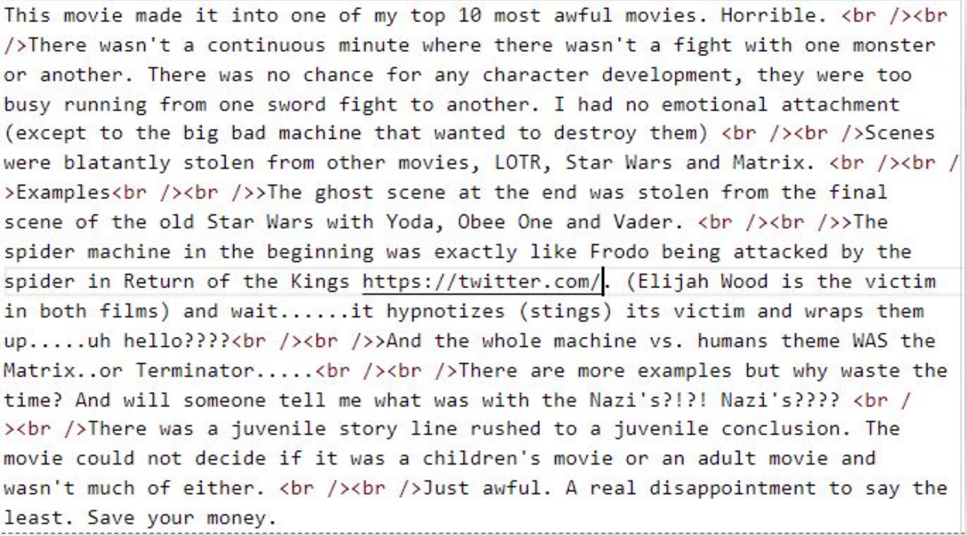


How are you?

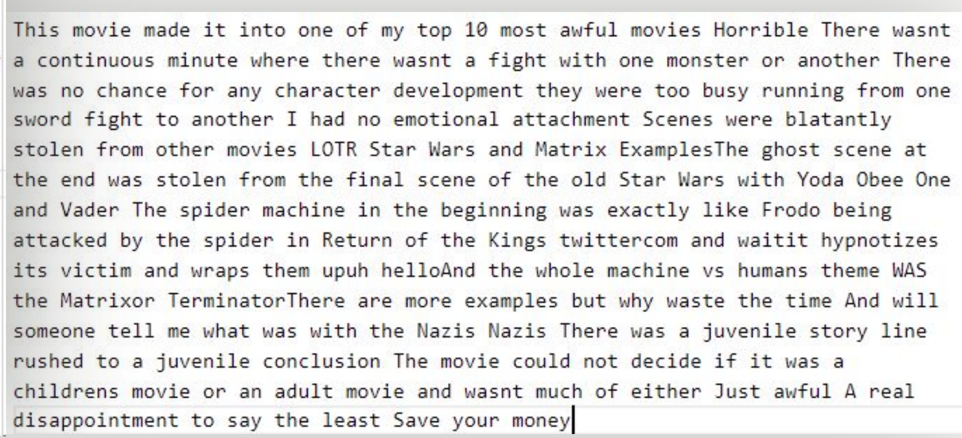
This is els test?

Special test:

Special harsens: is/



```
This movie made it into one of my top 10 most awful movies. Horrible. <br /><br
/>There wasn't a continuous minute where there wasn't a fight with one monster
or another. There was no chance for any character development, they were too
busy running from one sword fight to another. I had no emotional attachment
(except to the big bad machine that wanted to destroy them) <br /><br />Scenes
were blatantly stolen from other movies, LOTR, Star Wars and Matrix. <br /><br /
>Examples<br /><br />>The ghost scene at the end was stolen from the final
scene of the old Star Wars with Yoda, Obee One and Vader. <br /><br />>The
spider machine in the beginning was exactly like Frodo being attacked by the
spider in Return of the Kings https://twitter.com/. (Elijah Wood is the victim
in both films) and wait......it hypnotizes (stings) its victim and wraps them
up.....uh hello????<br /><br />>And the whole machine vs. humans theme WAS the
Matrix..or Terminator.....<br /><br />There are more examples but why waste the
time? And will someone tell me what was with the Nazi's?!?! Nazi's???? <br /
><br />There was a juvenile story line rushed to a juvenile conclusion. The
movie could not decide if it was a children's movie or an adult movie and
wasn't much of either. <br /><br />Just awful. A real disappointment to say the
least. Save your money.
```



```
This movie made it into one of my top 10 most awful movies Horrible There wasnt
a continuous minute where there wasnt a fight with one monster or another There
was no chance for any character development they were too busy running from one
sword fight to another I had no emotional attachment Scenes were blatantly
stolen from other movies LOTR Star Wars and Matrix ExamplesThe ghost scene at
the end was stolen from the final scene of the old Star Wars with Yoda Obee One
and Vader The spider machine in the beginning was exactly like Frodo being
attacked by the spider in Return of the Kings twittercom and waitit hypnotizes
its victim and wraps them upuh helloAnd the whole machine vs humans theme WAS
the Matrixor TerminatorThere are more examples but why waste the time And will
someone tell me what was with the Nazis Nazis There was a juvenile story line
rushed to a juvenile conclusion The movie could not decide if it was a
childrens movie or an adult movie and wasnt much of either Just awful A real
disappointment to say the least Save your money
```

Noise Removal: URLs and HTML Tags

1 Remove URLs

Strip web links to clean text content

2 Remove HTML tags

Eliminate markup language artifacts from raw text

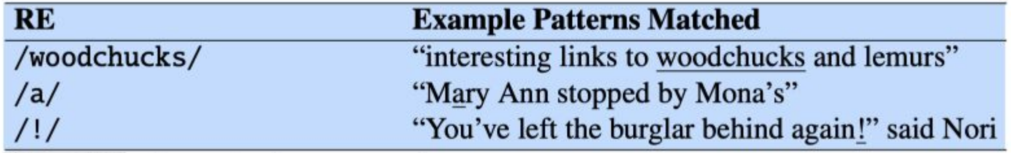
Example:

```
    re.findall(r'\b\d{10}\b', text)  # Extract phone numbers
```

Regular expression

- Regular expressions (REs) are a powerful tool for specifying patterns to search, match, or manipulate text strings.
- **Why Use Them in NLP?**
  - Quickly find words, phrases, or patterns in large text corpora
  - Essential for tasks like tokenization, data cleaning, and information extraction
- **Examples of Patterns:**
  - Email addresses: `\w+@\w+\.\w+`
  - Dates: `\d{2}/\d{2}/\d{4}`
  - Words ending with “ing”: `\b\w+ing\b`

Basic Regex

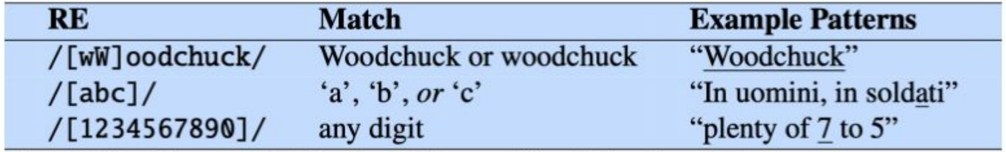


RE	Example Patterns Matched
/woodchucks/	“interesting links to woodchucks and lemurs”
/a/	“Mary Ann stopped by Mona’s”
/!/	“You’ve left the burglar behind again!” said Nori

Figure 2.1 Some simple regex searches.

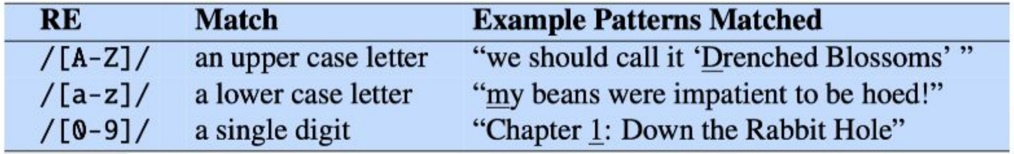
- Most characters match themselves
- Sequences of regexps match sequences of characters

- Brackets indicate a set of possible single-character matches



RE	Match	Example Patterns
/[wW]oodchuck/	Woodchuck or woodchuck	“Woodchuck”
/[abc]/	‘a’, ‘b’, *or* ‘c’	“In uomini, in soldati”
/[1234567890]/	any digit	“plenty of 7 to 5”

Figure 2.2 The use of the brackets [] to specify a disjunction of characters.

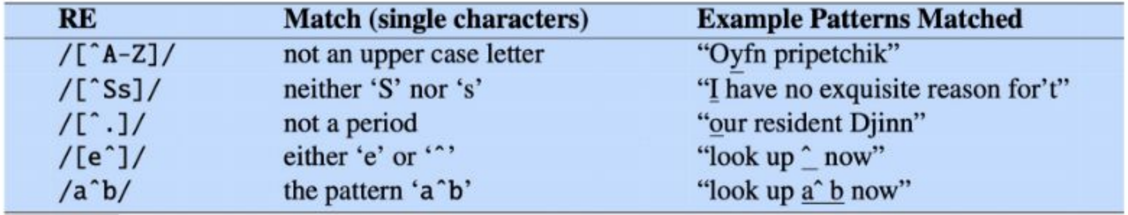


RE	Match	Example Patterns Matched
/[A-Z]/	an upper case letter	“we should call it ‘Drenched Blossoms’ ”
/[a-z]/	a lower case letter	“my beans were impatient to be hoed!”
/[0-9]/	a single digit	“Chapter 1: Down the Rabbit Hole”

Figure 2.3 The use of the brackets [] plus the dash - to specify a range.

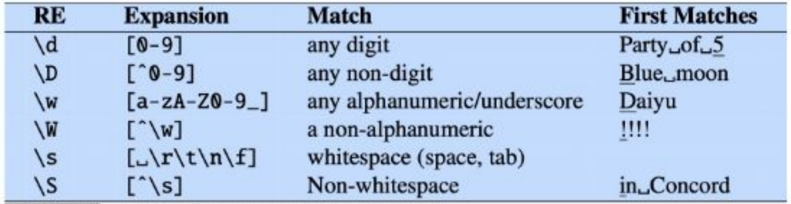
- A dash (hypen) inside a bracket implies a character range

- A caret (as the first character of the regexp) complements the set of characters in the range



RE	Match (single characters)	Example Patterns Matched
/[^A-Z]/	not an upper case letter	"Oyfn pripetchik"
/[^Ss]/	neither 'S' nor 's'	"I have no exquisite reason for't"
/[^.]/	not a period	"our resident Djinn"
/[e^]/	either 'e' or '^'	"look up ^ now"
/a^b/	the pattern 'a^b'	"look up a^ b now"

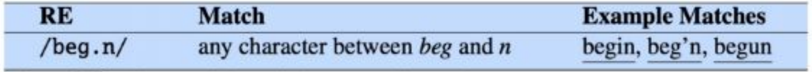
Figure 2.4 The caret ^ for negation or just to mean ^. See below re: the backslash for escaping the period.



RE	Expansion	Match	First Matches
\d	[0-9]	any digit	Party␣of␣5
\D	[^0-9]	any non-digit	Blue␣moon
\w	[a-zA-Z0-9_]	any alphanumeric/underscore	Daiyu
\W	[^\w]	a non-alphanumeric	!!!!
\s	[␣\r\t\n\f]	whitespace (space, tab)	
\S	[^\s]	Non-whitespace	in␣Concord

Figure 2.8 Aliases for common sets of characters.

- Named ranges use backslash notation for common patterns: digits, alphanumerics and whitespace

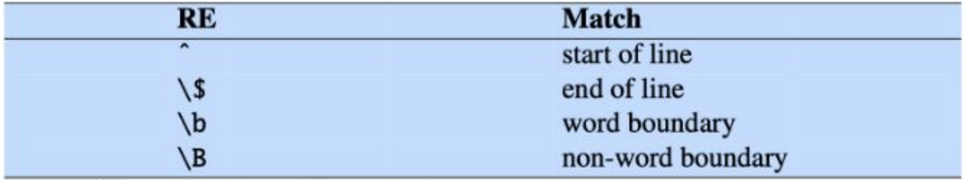


RE	Match	Example Matches
/beg.n/	any character between *beg* and *n*	begin, beg'n, begun

Figure 2.6 The use of the period . to specify any character.

- A dot matches any character (wildcard)

- Matches can be forced to be at the beginning or end of a word or line



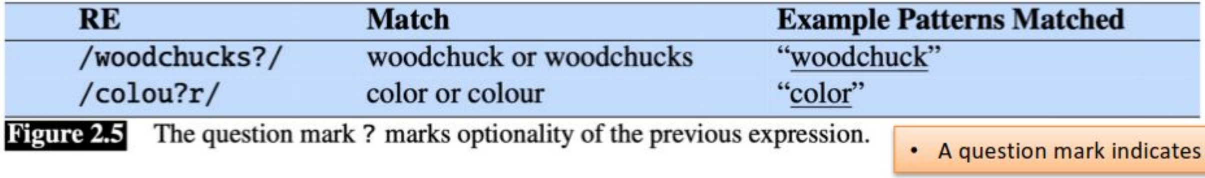
RE	Match
^	start of line
\$	end of line
\b	word boundary
\B	non-word boundary

Figure 2.7 Anchors in regular expressions.

Repetition

Compare:

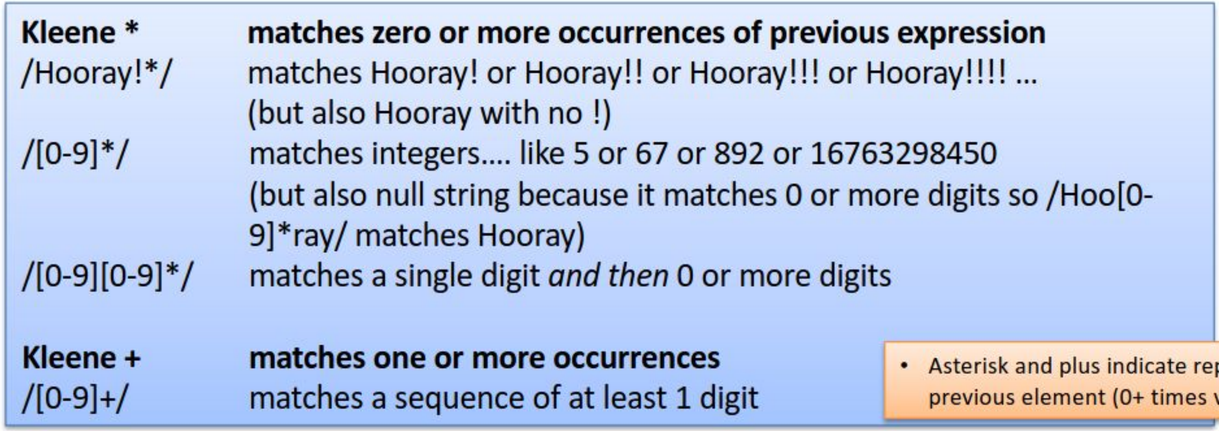
- re.search("a.*c", "abcabc").group()
- re.search("a.*?c", "abcabc").group()



RE	Match	Example Patterns Matched
/woodchucks?/	woodchuck or woodchucks	"woodchuck"
/colou?r/	color or colour	"color"

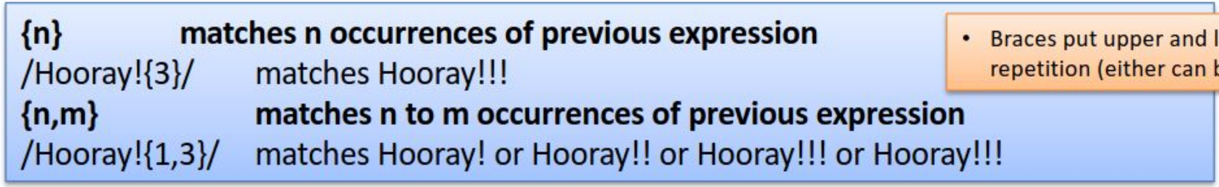
Figure 2.5 The question mark ? marks optionality of the previous expression.

- A question mark indicates optionality



Kleene *	**matches zero or more occurrences of previous expression**
/Hooray!*/	matches Hooray! or Hooray!! or Hooray!!! or Hooray!!!! … (but also Hooray with no !)
/[0-9]*/	matches integers…. like 5 or 67 or 892 or 16763298450 (but also null string because it matches 0 or more digits so /Hoo[0-9]*ray/ matches Hooray)
/[0-9][0-9]*/	matches a single digit *and then* 0 or more digits
Kleene +	**matches one or more occurrences**
/[0-9]+/	matches a sequence of at least 1 digit

- Asterisk and plus indicate repetition of the previous element (0+ times vs 1+ times)

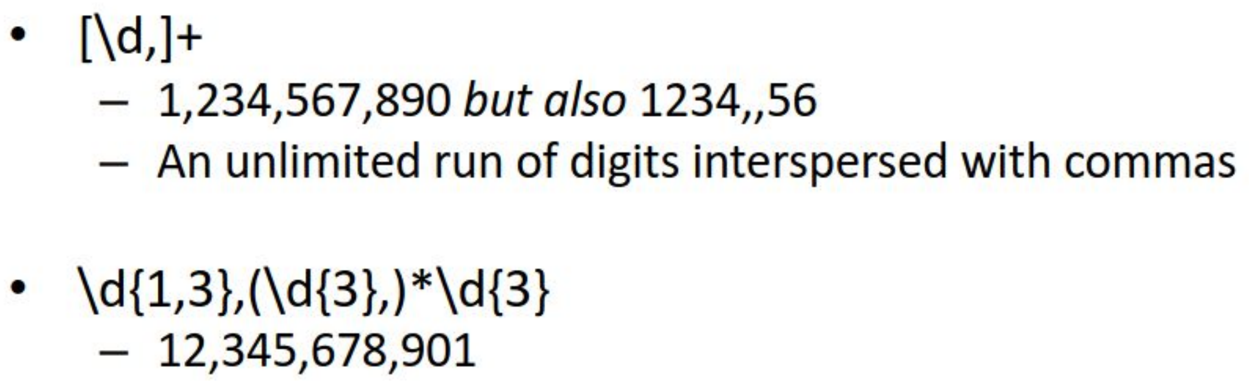


{n}	**matches n occurrences of previous expression**
/Hooray!{3}/	matches Hooray!!!
{n,m}	**matches n to m occurrences of previous expression**
/Hooray!{1,3}/	matches Hooray! or Hooray!! or Hooray!!! or Hooray!!!

- Braces put upper and lower bounds on repetition (either can be blank)

Extended Regex

- **Integer parsing**
- Alternatives and Groups
- Capture Groups

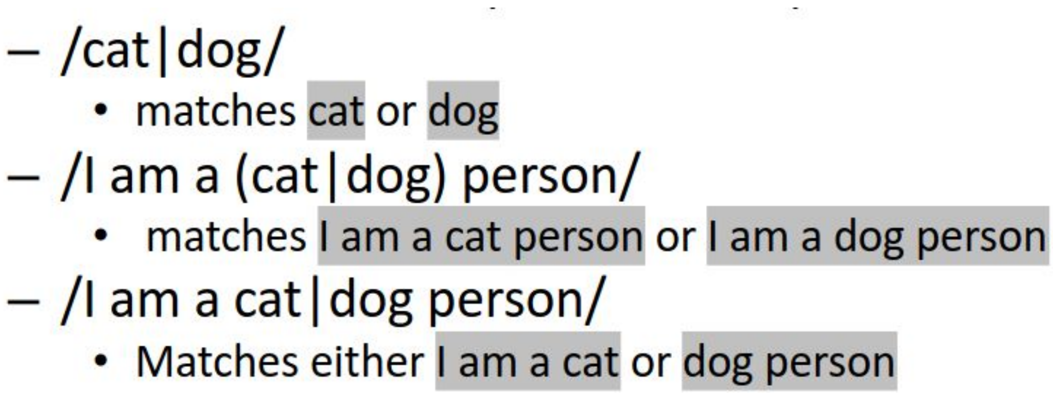


- [\d,]+
  - 1,234,567,890 *but also* 1234,,56
  - An unlimited run of digits interspersed with commas

- \d{1,3},(\d{3},)*\d{3}
  - 12,345,678,901

Extended Regex

- Integer parsing
- **Alternatives and Groups**
- Capture Groups

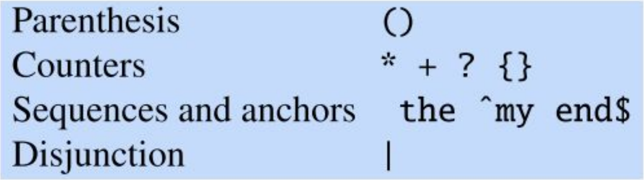


- /cat|dog/
  - matches cat or dog
- /I am a (cat|dog) person/
  - matches I am a cat person or I am a dog person
- /I am a cat|dog person/
  - Matches either I am a cat or dog person

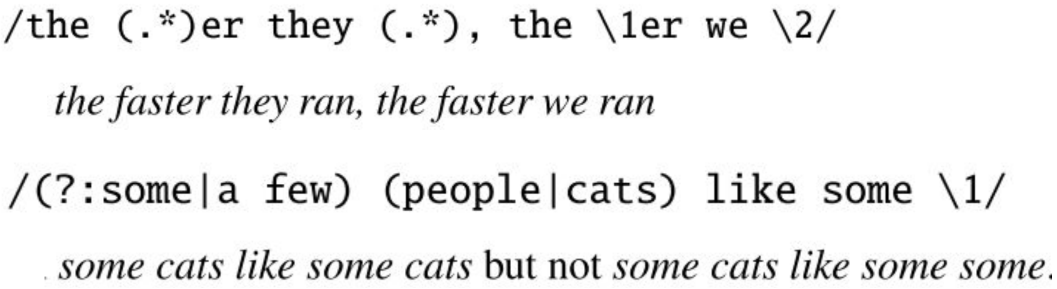
Extended Regex

- Integer parsing
- Alternatives and Groups
- **Capture Groups**

Operator precedence hierarchy



Parenthesis	`()`
Counters	`* + ? {}`
Sequences and anchors	`the ˆmy end$`
Disjunction	`|`



```
/the (.*)er they (.*), the \1er we \2/
```

the faster they ran, the faster we ran

```
/(?:some|a few) (people|cats) like some \1/
```

some cats like some cats but not *some cats like some some*.

non-capturing group

How Are Regular Expressions Used in NLP?

- **Common Applications:**
  - **Tokenization:** Splitting text into words, sentences, or tokens
  - **Data Validation:** Checking formats of emails, dates, URLs
  - **Text Filtering:** Removing stopwords, unwanted characters
  - **Pattern Extraction:** Identifying entities, abbreviations, or specific phrases
  - **Text Transformation:** Replacing or reformatting parts of text (e.g., converting “I’m” to “I am”)
- **Real-World Example:**
  - The classic chatbot ELIZA used a cascade of regular expression substitutions to understand and respond to user input:
    s/.* YOU ARE (depressed|sad) .*/I AM SORRY TO HEAR YOU ARE \1/

Quick recap of today’s lecture

Why text preprocessing is important?

- **Noise Reduction:** Remove irrelevant characters (punctuation, HTML tags).
- **Consistency:** Lowercasing, standardizing formats (e.g., dates).
- **Efficiency:** Smaller vocabulary size = faster model training.
- **Accuracy:** Improves NLP task performance (e.g., sentiment analysis).

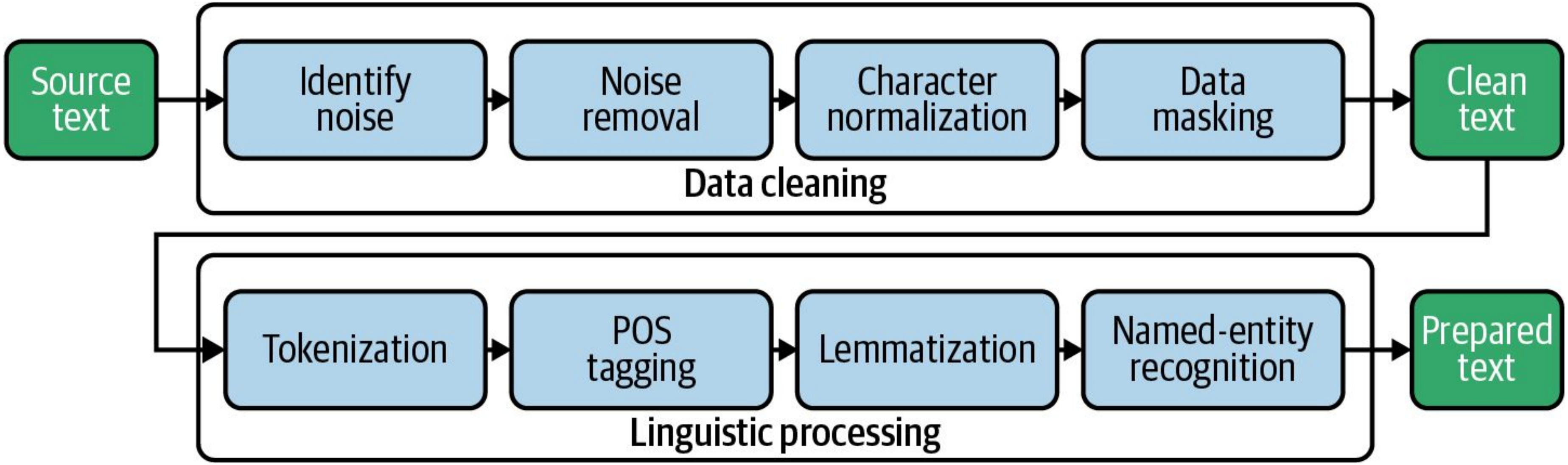
Challenges

- **Ambiguity:** “Apple” (fruit vs. company).
- **Language Differences:** Morphology in Arabic vs. English.
- **Resource Limits:** Lemmatization requires heavy dictionaries.

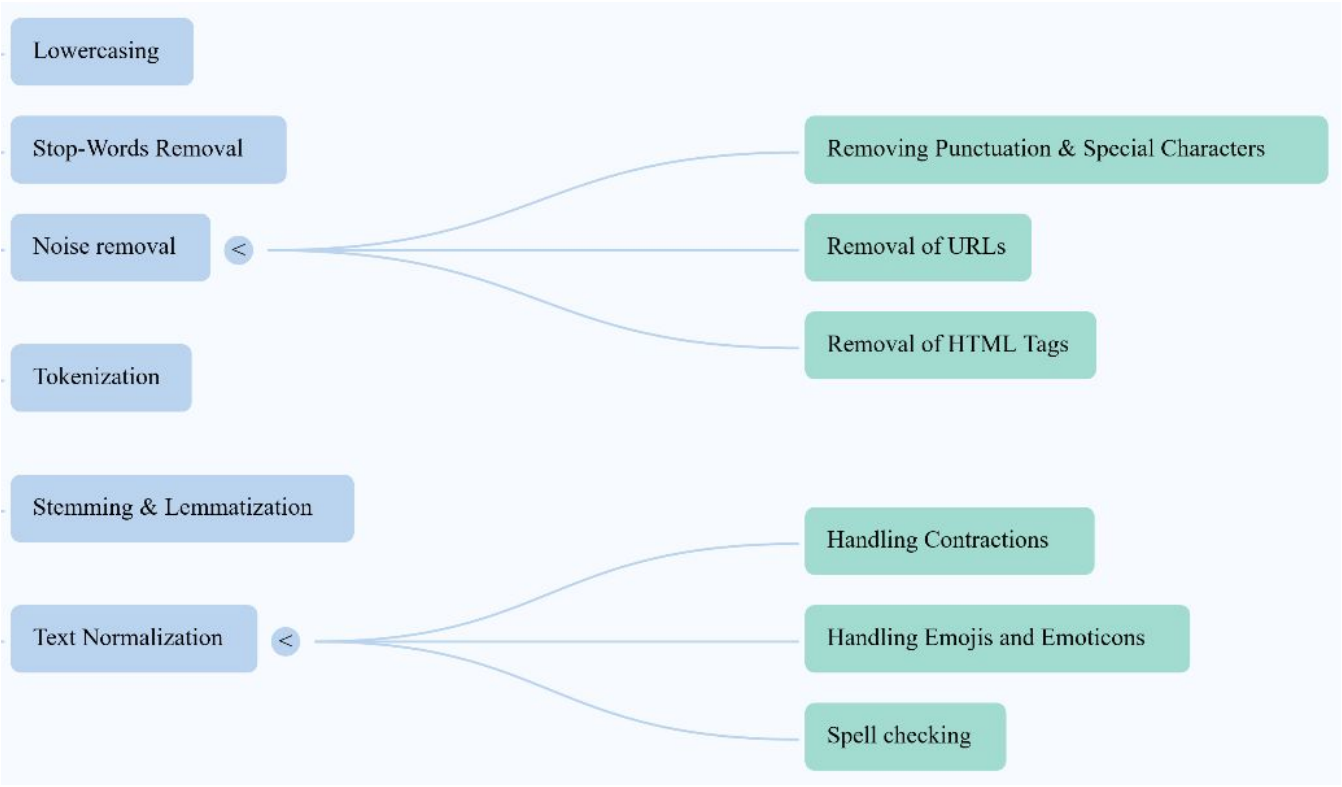
Domain-Specific Preprocessing

- **Social Media:** Emoji handling, slang normalization.
- **Scientific Texts:** Retain equations/symbols.
- **Medical Texts:** Protect sensitive terms (e.g., patient names).

Text preprocessing stages



Common preprocessing steps:



- Lowercasing
- Stop-Words Removal
- Noise removal
  - Removing Punctuation & Special Characters
  - Removal of URLs
  - Removal of HTML Tags
- Tokenization
- Stemming & Lemmatization
- Text Normalization
  - Handling Contractions
  - Handling Emojis and Emoticons
  - Spell checking

Text preprocessing techniques in NLP

- Regular expression
- Tokenization
- Lemmatization
- Stemming
- Part-of-speech (POS) tagging
- Name-Entity-Recognition (NER)

Toolkits:

- **NLTK:** Noise removal, Tokenization, stemming, POS/NER tagging.
- **spaCy:** Tokenization, stemming, POS/NER tagging, Dependency Parsing.
- **Hugging Face Tokenizers:** For transformer models.

References

- Term frequency (TF), Inverse Document Frequency (IDF), TF-IDF [https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/]
- Regex [Chapter 2: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf]

Practical applications:

- Generating Word Cloud in Python [https://www.geeksforgeeks.org/python/generating-word-cloud-python/]
  a. Visualize stopwords using wordcloud based on the score of TF, IDF, and TF-IDF.
- Removal of stopwords using NLP toolkits [https://www.geeksforgeeks.org/nlp/removing-stop-words-nltk-python/]
  a. Compare the stopwords filtered using TF, IDF, and TF-IDF based methods vs NLP toolkits.

* **Image Sources:** Educational use from Google Search and publicly available NLP materials.