# First-Year Project 3: Natural Language Processing

Christian Hardmeier

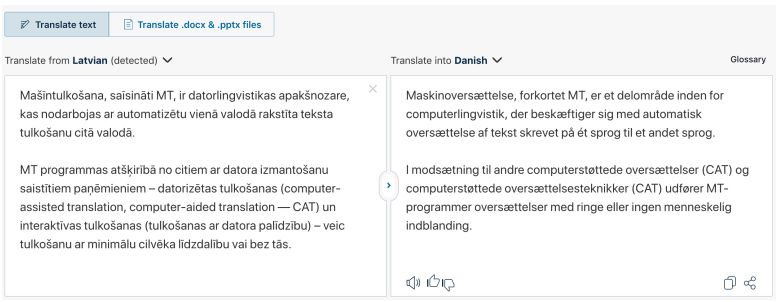
19 April 2022

IT UNIVERSITY OF COPENHAGEN

# Natural Language Processing

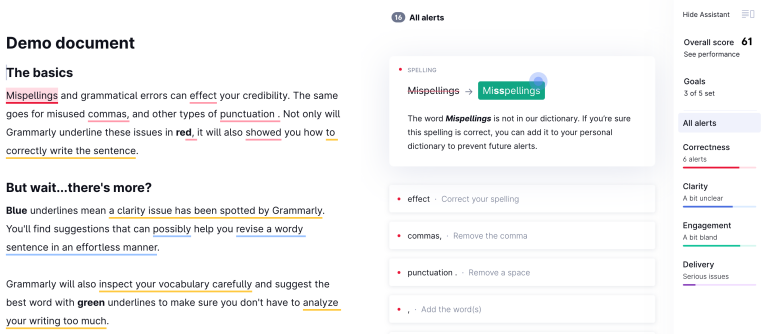
IT UNIVERSITY OF COPENHAGEN

## Machine Translation



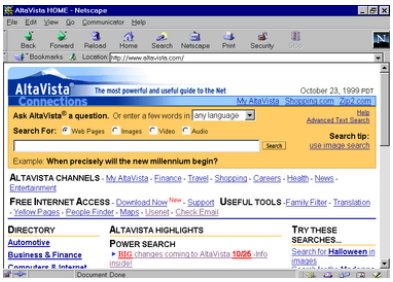

IT UNIVERSITY OF COPENHAGEN

# Writing Aids



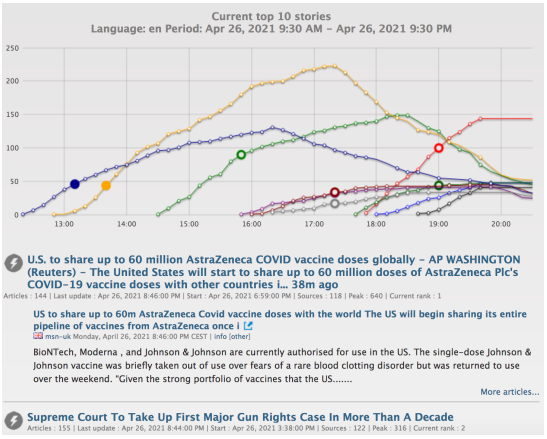

IT UNIVERSITY OF COPENHAGEN

# Information retrieval




IT UNIVERSITY OF COPENHAGEN

# Information aggregation/extraction




https://emm.newsbrief.eu/

IT UNIVERSITY OF COPENHAGEN

# Natural Language

IT UNIVERSITY OF COPENHAGEN

**Where Young College Graduates Are Choosing to Live**
*New York Times, 20 Oct 2014*

When young college graduates decide where to move, they are not just looking at the usual suspects, like New York, Washington and San Francisco. Other cities are increasing their share of these valuable residents at an even higher rate and have reached a high overall percentage, led by Denver, San Diego, Nashville, Salt Lake City and Portland, Ore., according to a report published Monday by City Observatory, a new think tank.

And as young people continue to spurn the suburbs for urban living, more of them are moving to the very heart of cities — even in economically troubled places like Buffalo and Cleveland. The number of college-educated people age 25 to 34 living within three miles of city centers has surged, up 37 percent since 2000, even as the total population of these neighborhoods has slightly shrunk.

Some cities are attracting young talent while their overall population falls, like Pittsburgh and New Orleans. And in a reversal, others that used to be magnets, like Atlanta and Charlotte, are struggling to attract them at the same rate.

Mol Pharm. 2014 Oct 15. [Epub ahead of print]

**Moxifloxacin Loaded Nanoemulsions having Tocopheryl Succinate as Integral Component Improves Pharmacokinetics and Enhances Survival in E Coli Induced Complicated Intra Abdominal Infection.**

Shukla P, Verma AK, Dwivedi P, Yadav A, Gupta PK, Rath SK, Mishra PR.

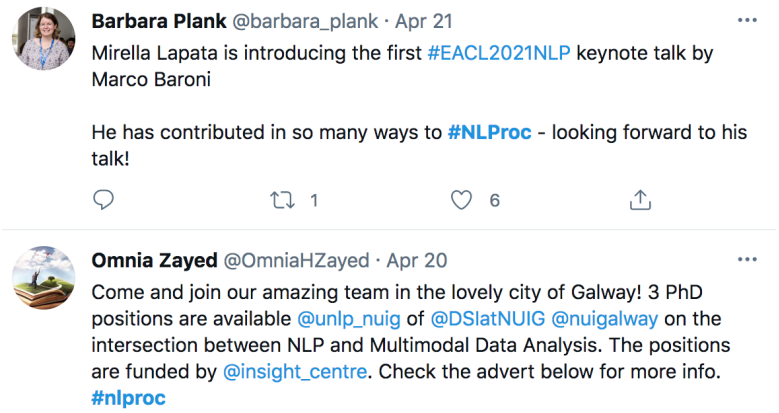
**Abstract**

In the present work a novel nanoemulsion laden with moxifloxacin has been developed for effective management of complicated Intra-abdominal infections. Moxifloxacin nanoemulsion fabricated using high pressure homogenization was evaluated for various pharmaceutical parameters, pharmacokinetics and pharmacodynamics in rats with E coli induced sepsis. The developed nanoemulsion MONe6 (size 168±28 d.nm and ZP -24.78±0.45mV respectively) was effective for intracellular delivery and sustaining the release of MOX. MONe6 demonstrated improved plasma (AUC MONe6/MOX= 2.38 fold) and tissue pharmacokinetics of MOX (AUC MONe6/MOX= 2.63 and 1.47 times in lung and **liver** respectively). Calculated PK/PD index correlated well with reduction in bacterial burden in plasma as well as tissues. Enhanced survival on treatment with MONe6 (65.44 %) and as compared to control group (8.22 %) was result of reduction in lipid peroxidation, neutrophil migration and cytokine levels (TNF-α and IL1b) as compared to untreated groups in rat model of E coli induced sepsis. Parenteral nanoemulsions of MOX hold a promising advantage in the therapy of E. coli induced complicated intra-abdominal infections and is helpful in the prevention of further complication like septic shock and death.

Athough people enjoyment differs from stage to stage, youth enjoys more as compared with the older people is definitely reasonable to agree with the statement. I observed in my life at part, that young people enjoys more.

In modern world, due to the development of technology, they are more fascinated to new kind of objects, such as mobiles, computer, internet, etc., and by using them that were available enjoying life.

Youngers are more affectionated to the day to day development of technology, by using those things they enjoying more. Older people are not that much aware of these technologies. Because technology is developing rapidly in these day to day life. Younger ones are more energetic than older people, they can go and do whatever they want. They make their friends as a part to enjoy in most of the public places, they usually attends party and meetings, and have capability to think and improve by adapting new technologies they enjoys more.






# What distinguishes natural language data?

## Tabular data

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | date | iso3166-2 | RelativeHumic | SolarRadiation | Surfacepressur | TemperatureA | Totalprecipitat | UVIndex | WindSpeed |
| 2 | 13/02/2020 | DE-BB | 76.337444 | 1824290.332 | 2403340.782 | 276.551573 | 0.003355 | 2.777806 | 4.542822 |
| 3 | 13/02/2020 | DE-BE | 76.065297 | 1786373.223 | 2408181.703 | 276.844633 | 0.003523 | 4.671329 | 4.761509 |
| 4 | 13/02/2020 | DE-BW | 80.113988 | 1505759.748 | 2290157.738 | 276.227143 | 0.008013 | 4.268546 | 4.467024 |
| 5 | 13/02/2020 | DE-BY | 81.554346 | 2363012.95 | 2275361.323 | 275.583053 | 0.005227 | 4.417797 | 3.677414 |
| 6 | 13/02/2020 | DE-HB | 87.167414 | 8389.755794 | 2406939.731 | 276.237452 | 0.007715 | 1.794872 | 4.699573 |
| 7 | 13/02/2020 | DE-HE | 89.200446 | 307067.7687 | 2331353.182 | 275.353974 | 0.005441 | 2.624676 | 4.398072 |
| 8 | 13/02/2020 | DE-HH | 86.685698 | 209488.6014 | 2404975.492 | 275.735476 | 0.005799 | 3.242424 | 4.363794 |
| 9 | 13/02/2020 | DE-MV | 81.246496 | 2115141.105 | 2407069.642 | 276.355814 | 0.002628 | 2.02164 | 4.910115 |
| 10 | 13/02/2020 | DE-NI | 87.179525 | 60634.84104 | 2391909.217 | 275.871953 | 0.007431 | 1.937668 | 4.307425 |
| 11 | 13/02/2020 | DE-NW | 85.525648 | 441980.9547 | 2361307.608 | 276.756529 | 0.007882 | 2.897228 | 4.687521 |
| 12 | 13/02/2020 | DE-RP | 89.927054 | 765036.3151 | 2326436.248 | 276.057197 | 0.006326 | 3.712812 | 4.820924 |
| 13 | 13/02/2020 | DE-SH | 88.33687 | 523403.7544 | 2406366.352 | 275.98947 | 0.004933 | 1.525916 | 4.786011 |
| 14 | 13/02/2020 | DE-SL | 88.784533 | 910951.3071 | 2330070.101 | 276.542441 | 0.00758 | 4.013587 | 5.792912 |
| 15 | 13/02/2020 | DE-SN | 74.823272 | 1352189.138 | 2340669.027 | 275.849612 | 0.002521 | 3.715195 | 5.085935 |
| 16 | 13/02/2020 | DE-ST | 80.708839 | 416147.9026 | 2385273.835 | 276.085706 | 0.004409 | 2.103347 | 4.396736 |
| 17 | 13/02/2020 | DE-TH | 83.149783 | 181806.4619 | 2317997.488 | 275.243508 | 0.004016 | 3.009039 | 5.137631 |
| 18 | 13/02/2020 | DK-81 | 86.342782 | 2555655.411 | 2403425.986 | 276.295413 | 0.000464 | 2.071701 | 3.384652 |
| 19 | 13/02/2020 | DK-82 | 87.751343 | 2153148.911 | 2399142.414 | 276.727344 | 0.001325 | 2.990025 | 3.801099 |
| 20 | 13/02/2020 | DK-83 | 86.736915 | 1451332.97 | 2402262.757 | 276.874953 | 0.003497 | 3.131156 | 4.773278 |
| 21 | 13/02/2020 | DK-84 | 84.394247 | 3058965.136 | 2409240.225 | 277.120543 | 0.00129 | 2.842956 | 4.884642 |
| 22 | 13/02/2020 | DK-85 | 86.200033 | 2891116.894 | 2408930.362 | 276.913832 | 0.002166 | 3.071611 | 5.276005 |
| 23 | 13/02/2020 | NL-DR | 91.711788 | 106694.1643 | 2403207.339 | 276.92834 | 0.010772 | 0.26178 | 4.431883 |
| 24 | 13/02/2020 | NL-FL | 89.725649 | 605101.3219 | 2402641.712 | 278.164065 | 0.012001 | 0.785507 | 6.057862 |
| 25 | 13/02/2020 | NL-FR | 91.979761 | 338643.5391 | 2402547.396 | 277.471187 | 0.009652 | 0.898502 | 5.811592 |



## Structured representation

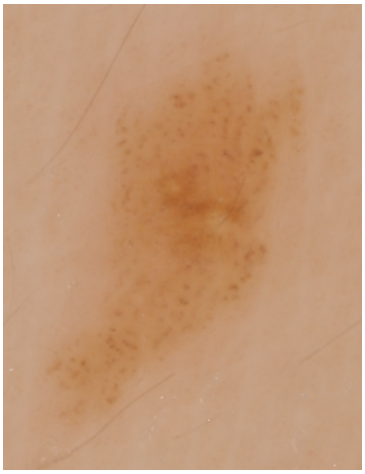
- Set of **data points** with associated **attributes** (or features).
- Known **metadata** describing the format and meaning of each attribute:

| | | |
|---|---|---|
| *Date* | Day/Month/Year | Gregorian calendar |
| *ISO3166-2* | Alphanumeric code | Closed list |
| *Relative humidity* | Fractional number | Percent |
| *Solar radiation* | Fractional number | ??? |
| *Surface pressure* | Fractional number | $24 \times \mu$bar ??? |
| *Temperature* | Fractional number | Kelvin |

- Observations are **interpreted** in a particular way.



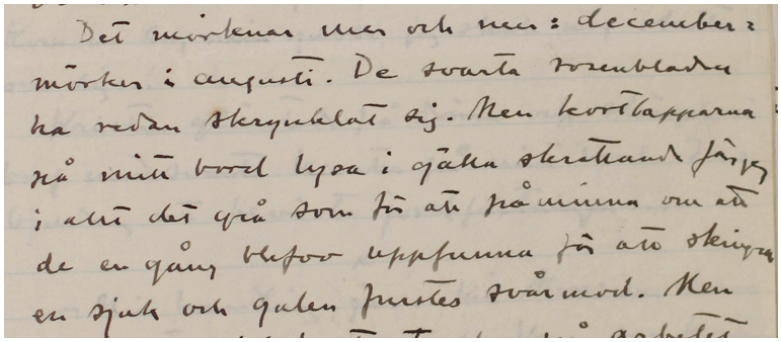
## Image data



# Unstructured 2D signal

- 2-dimensional spatial organisation.
  Every pixel has 8 neighbours.
- Each individual pixel carries almost no information.
  Adding noise to any or all pixels doesn't change the picture.
- Structure and meaning arise from pixels in context.
- Strong correlations between nearby elements.





IT UNIVERSITY OF COPENHAGEN

på svenska. Det är på det hela taget ett förbannat språk vi ha. Orden trampa varandra på tårna och knuffa varandra i rännstenen. Och allt blir så påtagligt och rått. Inga halvtoner, inga lätta antydningar och mjuka övergångar. Ett språk som tycks vara skapat till bruk för den outrotliga pöbelvanan att plumpa ut med sanningen i alla väder.

Det mörknar mer och mer: decembermörker i augusti. De svarta rosenbladen ha redan skrynklat sig. Men kortlapparna på mitt bord lysa i gälla skrattande färger i allt det grå som för att påminna om att de en gång blevo uppfunna för att skingra en sjuk och galen furstes svårmod. Men jag fasar vid blotta tanken på arbetet att samla ihop dem och vända de aviga rätt och blanda dem till en ny patiens, jag kan bara sitta och se på dem och lyssna till hur "hjärterknekt och spaderdam viska dystert om sin begravda kärlek", som det står i samma sonett.

Le beau valet de cœur et la dame de pique
causent sinistrement de leurs amours défunts.

Jag kunde ha lust att gå upp i det smutsiga gamla rucklet där snett över och dricka porter med flickorna. Röka en sur pipa och dra en spader med värdinnan och ge henne goda råd för hennes reumatism. Hon var här i förra veckan och klagade sin nöd, fet och präktig. Hon hade en tjock guldbrosch under isterhakan och betalade en femma kontant. Hon skulle bli smickrad av en kontravisit.

158

Det ringde på tamburdörren. Nu öppnar Kristin . . . Vad kan det vara? Jag har ju sagt till att jag inte tar emot i dag . . . En detektiv? . . . Som låtsas vara sjuk, uppträder som patient . . . Kom in du, min gubbe, jag skall nog sköta om dig . . .

Kristin gläntade på dörren och slängde ett brev med svarta kanter på mitt bord. Inbjudning att bevista jordfästningen . . .

*

— Min handling, ja . . . "Vil Monsieur have den Historie paa heroiske Vers, saa koster det 8 Skilling . . ."

25 augusti.

Jag såg i drömmen gestalter från min ungdom. Jag såg henne som jag kysste en midsommarnatt för länge sedan, då jag var ung och icke hade dödat någon. Jag såg också andra unga flickor av dem som hörde till vår krets den tiden; en som gick och läste det året jag blev student och som alltid ville tala med mig om religionen; en annan, som var äldre än jag och som gärna stod och viskade med mig i skymningen bakom en jasminhäck i vår trädgård. Och en annan, som alltid gjorde narr av mig, men som blev så ond och häftig och föll i krampgråt en gång då jag gjorde narr av henne . . . De gingo bleka i en blek skymning, deras ögon stodo vidöppna och förskrämda, och de gjorde tecken åt

159

```
på svenska. Det är på det hela taget ett förbannat
språk vi ha. Orden trampa varandra på tårna och
knuffa varandra i rännstenen. Och allt blir så
påtagligt och rått. Inga halvtoner, inga lätta antydningar
och mjuka övergångar. Ett språk som tycks vara
skapat till bruk för den outrotliga pöbelvänan att
plumpa ut med sanningen i alla väder.

Det mörknar mer och mer: decembermörker i
augusti. De svarta rosenbladen ha redan skrynklat
sig. Men kortlapparna på mitt bord lysa i gälla
skrattande färger i allt det grå som för att påminna
om att de en gång blevo uppfunna för att skingra
en sjuk och galen furstes svårmod. Men jag fasar
vid blotta tanken på arbetet att samla ihop dem
...
```



## Written text: Unstructured 1D signal with discrete elements

- One-dimensional spatial (linear) organisation.
  Every token has 2 neighbours.
- Discrete elements:
  Changing a word is *categorical*, not gradual.
- Hierarchical structure at multiple levels.
  Structural details poorly understood
  (competing linguistic theories).
- Strong correlations between nearby elements.
- Individual points (words/tokens) inherently meaningful,
  but ambiguous.



# Project Overview

## Overview

- Social media data from Twitter
- One tweet per line. User handles anonymised as *@user*.
- Predict speaker's *intentions* or *state of mind* (pragmatics).
- Supervised classification:
  - Given an input, predict a label.
  - Trained on data with manually annotated labels.
- 7 seven different tasks – pick 2 of them!
- *Binary* classification: 2 classes
  *Multiclass* classification: more than 2 classes



## Binary classification tasks

- Irony Detection
  - *Off to bed can't wait to feel this hangover.*
  - *Jeez it's a lovely morning out!!* 💨💧💦 👎 *#Ireland #December*
- Offensive Language Detection
  - *@user @user @user She is a walking talking lie.. that's why.*
  - *@user Eric holder is Obama's straw man of corruption*
- Hate Speech Detection
  - Lots of really disgusting hate speech
  - Personal attacks, often sexualised



## Emotion Recognition

- ANGER
  *Ppl like that irritate my soul*
- JOY
  *Happy Birthday @user #cheer #cheerchick #jeep #jeepgirl #IDriveAJeep #jeepjeep #Cheer*
- OPTIMISM
  *The point of living, and being an optimist, is to be foolish enough to believe the best is yet to come' - Peter Ustinov #optimism #quote*
- SADNESS
  *@user wow I'm just really sadden by that. Terrible*

## Emoji Prediction

20 labels – 😍, 😂, . . . , 🎄, 📷, 😜

- 😂
  *Man these are the funniest kids ever!! That face! #HappyBirthdayBubb @ FLIPnOUT Xtreme*
- 😍
  *Sundays are all about the cute babies and dogs! #Ballard #sundaymarket #littelestmodel. . .*
- 🎄
  *Christmas is up!! (@ The Dog House in Seattle, WA)*
- 🔥
  *A spicy Volcano Roll just erupted in my mouth! Delectable!*



## Sentiment Analysis

- NEGATIVE
  *Thanks manager for putting me on the schedule for Sunday*
- NEUTRAL
  *Who wants to be my date to the White Sox vs Red Sox game Tuesday*
- POSITIVE
  *Happy Birthday Nick J  May you live long and Happy :)*



## Stance Detection

3 labels – *favour, neutral, against*
in relation to 5 target topics:

- Abortion
- Atheism
- Climate change
- Feminism
- Hillary Clinton

## Stance Detection: Atheism

- NONE
  *@user Old age has not made you any wiser or more mature. For shame!*
- FAVOUR
  *If you regularly base your thoughts on superstitions, you might not be able to think well. #freethinker*
- AGAINST
  *Daily time in God's Word yields lasting freedom. #assurance*



## Where do those labels come from?

- Manually labelled (annotated) data: 6 out of 7 tasks
  - Very popular method to encode human knowledge in small to medium datasets
  - Requires clear task definition and guidelines
  - Quality control with multiple annotators
- *Fortuitous* data: Emoji classification

*@user it's all good* 😘 ⟨ TEXT: *@user it's all good*; LABEL: 😘 ⟩



## What will we learn?

- Basic properties of natural language
- Preprocessing: How to prepare texts for automatic processing?
- Annotation:
  - How to label up data for supervised classification?
  - How to evaluate the quality of manual annotations?
- Basic NLP feature extraction and classification
- Basic language modelling and data augmentation
- Evaluation of classifiers in NLP

# Preprocessing

IT UNIVERSITY OF COPENHAGEN

**Data acquisition**

From a customer/ employer

From a data provider (LDC, ELRA, ELRC)

Web scraping

Human informants

**Data formats**

Presentation-oriented formats (HTML, DOC/DOCX, PDF)

Encodings (“code pages”; Unicode, Latin-1, . . . )

Metadata conservation:

*Where does the data come from?*

*Who produced it?*

*Where can I find more context?*

**Cleaning**

Language identification

Page headers/footers

Spelling errors

**Segmentation**

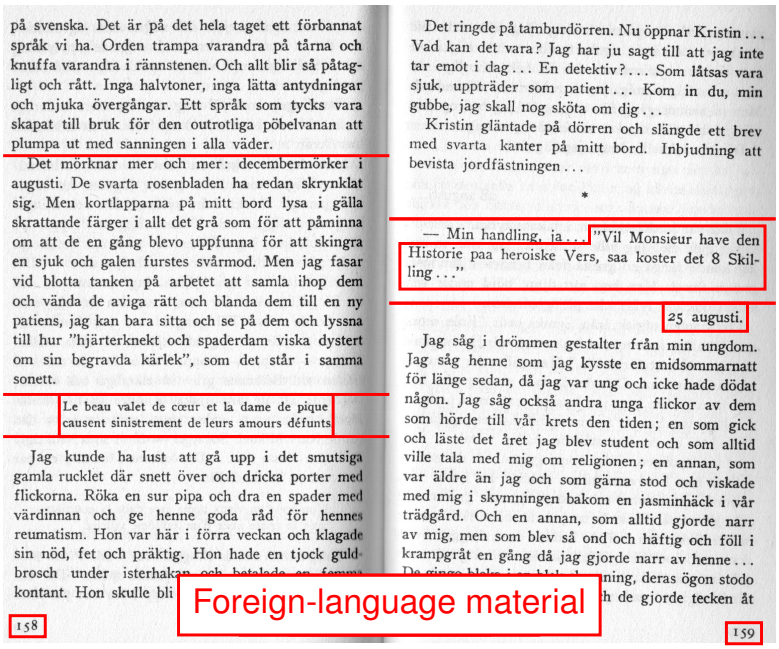
Documents

Chapters/sections

Sentences

Words

IT UNIVERSITY OF COPENHAGEN



på svenska. Det är på det hela taget ett förbannat språk vi ha. Orden trampa varandra på tårna och knuffa varandra i rännstenen. Och allt blir så påtagligt och rått. Inga halvtoner, inga lätta antydningar och mjuka övergångar. Ett språk som tycks vara skapat till bruk för den outrotliga pöbelvanan att plumpa ut med sanningen i alla väder.

Det mörknar mer och mer: decembermörker i augusti. De svarta rosenbladen ha redan skrynklat sig. Men kortlapparna på mitt bord lysa i gälla skrattande färger i allt det grå som för att påminna om att de en gång blevo uppfunna för att skingra en sjuk och galen furstes svårmod. Men jag fasar vid blotta tanken på arbetet att samla ihop dem och vända de aviga rätt och blanda dem till en ny patiens, jag kan bara sitta och se på dem och lyssna till hur ”hjärterknekt och spaderdam viska dystert om sin begravda kärlek”, som det står i samma sonett.

Le beau valet de cœur et la dame de pique causent sinistrement de leurs amours défunts

Jag kunde ha lust att gå upp i det smutsiga gamla rucklet där snett över och dricka porter med flickorna. Röka en sur pipa och dra en spader med värdinnan och ge henne goda råd för hennes reumatism. Hon var här i förra veckan och klagade sin nöd, fet och präktig. Hon hade en tjock guldbrosch under isterhakan och betalade en femma kontant. Hon skulle bli

158

Det ringde på tamburdörren. Nu öppnar Kristin . . . Vad kan det vara? Jag har ju sagt till att jag inte tar emot i dag . . . En detektiv? . . . Som låtsas vara sjuk, uppträder som patient . . . Kom in du, min gubbe, jag skall nog sköta om dig . . .

Kristin gläntade på dörren och slängde ett brev med svarta kanter på mitt bord. Inbjudning att bevista jordfästningen . . .

*

— Min handling, ja . . . ”Vil Monsieur have den Historie paa heroiske Vers, saa koster det 8 Skilling . . .”

25 augusti.

Jag såg i drömmen gestalter från min ungdom. Jag såg henne som jag kysste en midsommarnatt för länge sedan, då jag var ung och icke hade dödat någon. Jag såg också andra unga flickor av dem som hörde till vår krets den tiden; en som gick och läste det året jag blev student och som alltid ville tala med mig om religionen; en annan, som var äldre än jag och som gärna stod och viskade med mig i skymningen bakom en jasminhäck i vår trädgård. Och en annan, som alltid gjorde narr av mig, men som blev så ond och häftig och föll i krampgråt en gång då jag gjorde narr av henne . . .

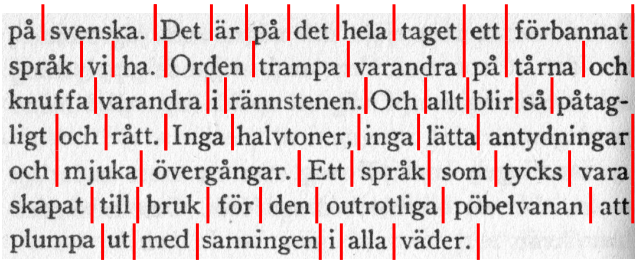
ning, deras ögon stodo

h de gjorde tecken åt

159

Foreign-language material

## Fine-grained segmentation




Word/token boundaries



## But... does that make sense?

svenska.

påtag-

ligt

halvtoner,

övergångar.



## Tokenisation

- *Tokenisation* or *word segmentation* is the task of splitting a text into minimal processing units
  - for further linguistic analysis
  - as input to machine learning solutions
- The units should be *meaningful* for the model or code that processes them.
- Models learn relations between tokens, so the segmentation should produce units that have meaningful relations.
- Often, “words” are a good level of segmentation to aim at.

## What is a word?

"the smallest meaningful sequence of sounds
that can be uttered in isolation"

- Definition from spoken language.
- In written language, we often assume that words are delimited by spaces or punctuation.
- The devil is in the detail:
  - *don't* –
  - *I'm* –
  - *dishwasher* –
  - *washing machine* –



## Reasons to prefer a particular segmentation

- Matching existing linguistic theory/annotation
- Matching existing models
- Trade-off between
  - Meaningfulness
  - Frequency
  - Number of parameters in models to train



# So how to do word segmentation?

## Regular expressions

- “Language” for pattern matching in texts.
- Efficient and expressive.
- Available in many programming languages and tools.
- Part of the Python standard library:

```
import re
```

- Basic principle:
  A regex *pattern matches* a string (or a part of it).



## Basic patterns

A␣cat␣sat␣on␣the␣mat.␣Next␣to␣it␣sat␣a␣rat.

- A letter (or any symbol without special meaning) matches itself.

  a t A at ␣
- Square brackets: Match one of the characters.

  [cs] [.␣]
- Negated square brackets: Match anything but those characters

  [^a-z] *(beware: äöå etc.!)*



## Special elements

- ^, $ – beginning and end of line
- \s – whitespace
- \S – not whitespace
- \w – word-internal character
- \W – not word-internal character
- and more: see library documentation

**Note:** \w uses a specific definition of *word*,
which may or may not be what you want!

## Combining elements

- *a** – *zero or more* occurrences of the previous item *a* (Kleene star)
- *a*+ – *one or more* occurrences of the previous item *a*
- *a*|*b* – alternatives
- (*abc*) – grouping, so you can write (*abc*)* or (*ab*|*c*)+



## Regex for string splitting

```
>>> import re
>>> line = 'A cat sat on the mat. His name was Måns.'
```

- If you can easily describe the *delimiter*: `re.split`

```
>>> re.split(' ', line)
['A', 'cat', 'sat', 'on', 'the', 'mat.', 'His', 'name', 'was', 'Måns.']
```

- If you can easily describe the *tokens*: `re.findall`

```
>>> re.findall(r'\w+', line)
['A', 'cat', 'sat', 'on', 'the', 'mat', 'His', 'name', 'was', 'Måns']
```

- Find a match *at the beginning of the string*: `re.match`
- Find a match *somewhere in the string*: `re.search`



## Separate patterns for tokens and delimiters

*code example*

## Tips for designing a tokeniser

- Start with a smallish subset of your data
  and tokenise it manually.
- Design regular expressions to match the desired tokenisation.
- Run it over a larger set, identify problems and refine.
- Looking through vocabulary lists can help you find problems
  (especially tokens that only occur once or twice).



## Intrinsic evaluation

Evaluating tokeniser “as tokeniser”.

- Requires a correct *gold standard segmentation*.
- *Word boundary recall:*

$$R = \frac{|C \cap G|}{|G|}$$

$R$: word boundary recall
$C$: word boundaries found by tokeniser
$G$: word boundaries in gold standard



## Extrinsic evaluation

Measure how the performance of a downstream task
depends on tokenisation.

- In the last part of the project, you will build a classifier.
- This allows you to do an extrinsic evaluation of the tokeniser.
- Wait until the end. . .

## Comparing tokenisations

- You can *compare* your tokenisation to that of another implementation and assess it qualitatively.
- Use the right tools to compare large outputs efficiently:
  - `difflib.SequenceMatcher` in the Python standard library
  - `diff` utility on Unix command line
- Your TAs can help you get going!



## Exercise

- Work on creating the tokenisation of your datasets.
  (Section 1 in the project description)
- Tokenisation is a prerequisite for everything that comes after!
- Your TAs are here to help you! Make good use of that!