TRACER TUTORIAL: TEXT REUSE DETECTION

PREPROCESSING

Marco Büchler, Emily Franzini and Greta Franzini

TRACER
TEXT REUSE DETECTION MACHINE

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

1. What is preprocessing?

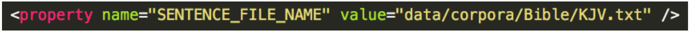
2. Preprocessing techniques

3. Hacking

4. Conclusion and revision

HACKING, INSTALLATION & CONFIGURATION GUIDE FOR TRACER

1. Download **TRACER** from `http://etrap.eu/tracer/` to your storage folder, e.g. `/roedel/mbuechler`
2. Using the command line, navigate to your storage folder with the `cd` command
3. **Unzip** archive: `gunzip tracer.tar.gz`
4. **Untar** archive: `tar -xvf tracer.tar`
5. Change to the TRACER folder: `cd TRACER`
6. **Open the configuration file** with `vim conf/tracer_config.xml`
7. Configure your input file:



```
<property name="SENTENCE_FILE_NAME" value="data/corpora/Bible/KJV.txt" />
```

HACKING: STARTING TRACER

Start the tool with the command:

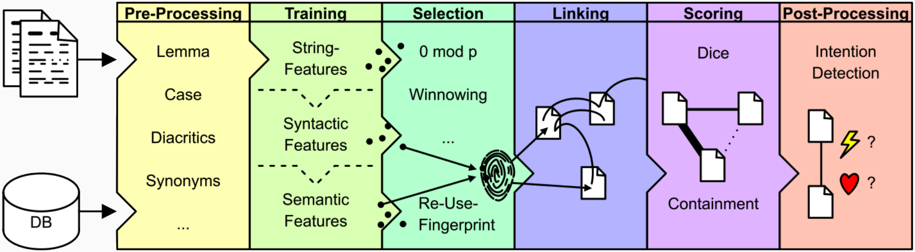
```
java -Xmx600m
-Deu.etrap.medusa.config.ClassConfig=conf/tracer_config.xml
-jar tracer.jar
```

Explanation:

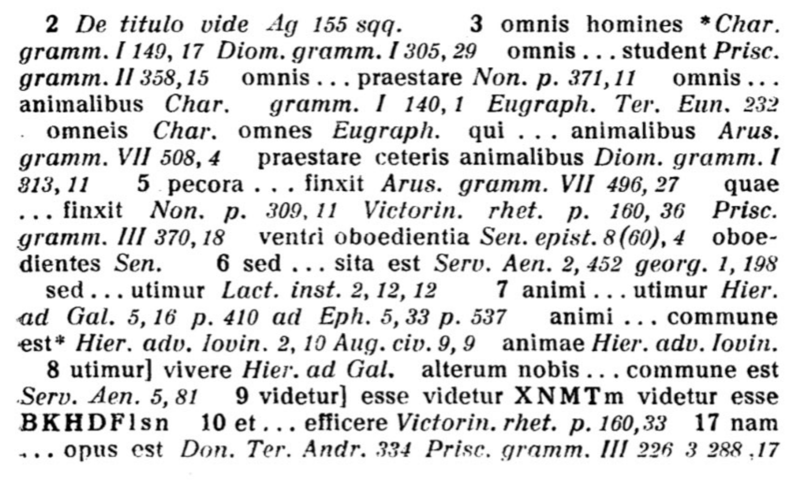
- `-Xmx600m` (up to 600 MB memory);
- `-Dfile.encoding` sets the encoding of your input file (optionally);
- `-Deu.etrap.medusa.config.ClassConfig` (configuration file).

WHAT IS PREPROCESSING?

REMINDER: CURRENT APPROACH

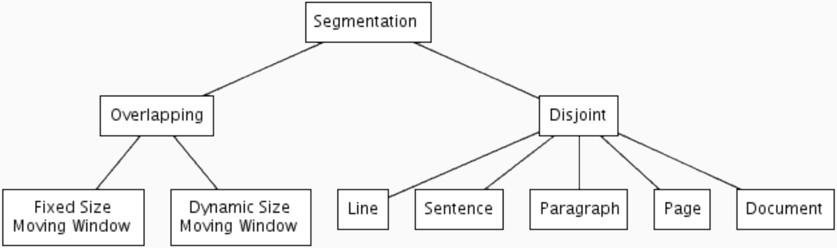


PRE-STEP: SEGMENTATION - AN EXAMPLE



2 *De titulo vide Ag 155 sqq.* 3 omnis homines **Char. gramm. I 149, 17 Diom. gramm. I 305, 29* omnis . . . student *Prisc. gramm. II 358, 15* omnis . . . praestare *Non. p. 371, 11* omnis . . . animalibus *Char. gramm. I 140, 1 Eugraph. Ter. Eun. 232* omneis *Char.* omnes *Eugraph.* qui . . . animalibus *Arus. gramm. VII 508, 4* praestare ceteris animalibus *Diom. gramm. I 313, 11* 5 pecora . . . finxit *Arus. gramm. VII 496, 27* quae . . . finxit *Non. p. 309, 11 Victorin. rhet. p. 160, 36 Prisc. gramm. III 370, 18* ventri oboedientia *Sen. epist. 8 (60), 4* oboedientes *Sen.* 6 sed . . . sita est *Serv. Aen. 2, 452 georg. 1, 198* sed . . . utimur *Lact. inst. 2, 12, 12* 7 animi . . . utimur *Hier. ad Gal. 5, 16 p. 410 ad Eph. 5, 33 p. 537* animi . . . commune est* *Hier. adv. Iovin. 2, 10 Aug. civ. 9, 9* animae *Hier. adv. Iovin.* 8 utimur] vivere *Hier. ad Gal.* alterum nobis . . . commune est *Serv. Aen. 5, 81* 9 videtur] esse videtur **XNMTm** videtur esse **BKHDFlsn** 10 et . . . efficere *Victorin. rhet. p. 160,33* 17 nam . . . opus est *Don. Ter. Andr. 334 Prisc. gramm. III 226 3 288 ,17*

PRE-STEP: SEGMENTATION



Segmentation

Overlapping

Disjoint

Fixed Size Moving Window

Dynamic Size Moving Window

Line

Sentence

Paragraph

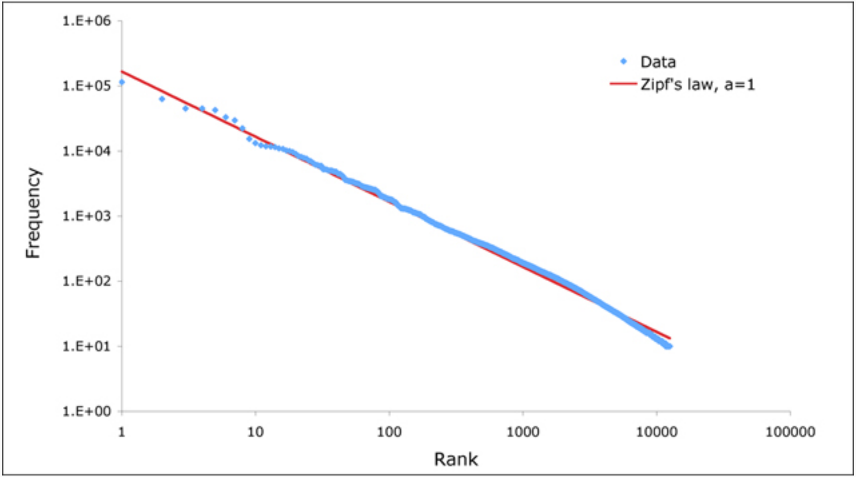
Page

Document

QUESTION

What do you associate with **preprocessing**?

FOUNDATIONS FOR PREPROCESSING: ZIPFIAN LAW



IMPLICATIONS OF THE ZIPFIAN LAW

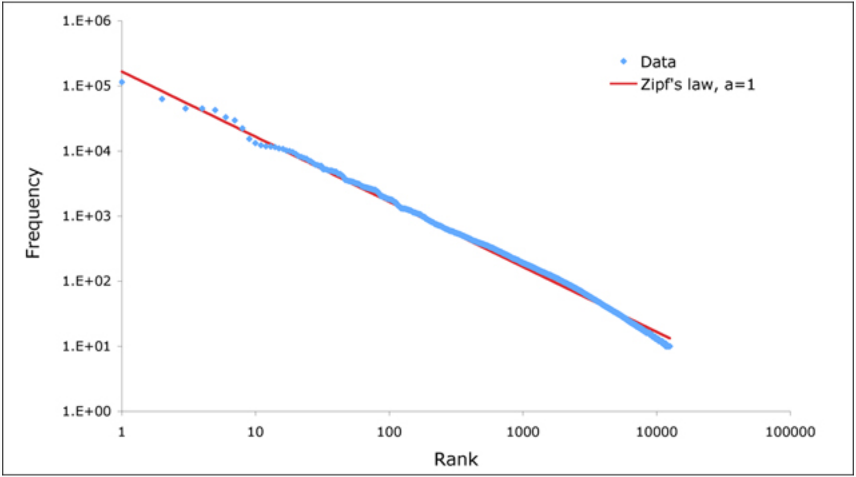
- **Approx. 50%** of all words occur only once
- **Approx. 16%** of all words occur only twice
- **Approx. 8%** of all words occur three times
- **...**
- **Approx. 90%** of all words in a corpus occur 10 times or less

$$s(f) = \frac{1}{f * (f+1)} \qquad s^n(f) = \sum_{f=1}^{n} \frac{1}{f * (f+1)}$$

- The **top 300-700 most** frequent words cover already about **50%** of all tokens (depending language)

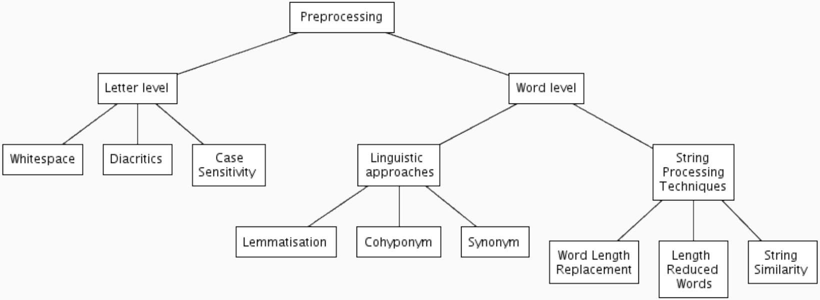
QUESTION

What does lemmatisation mean for this plot?



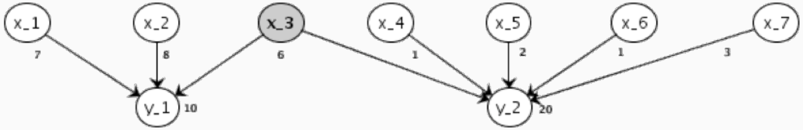
PREPROCESSING TECHNIQUES

PREPROCESSING



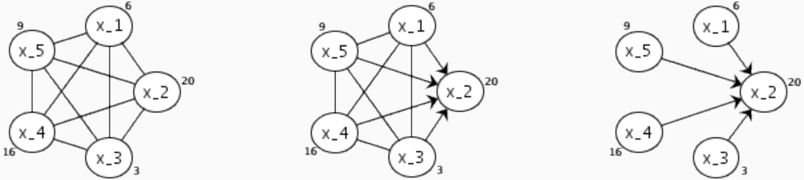
PREPROCESSING: DIRECTED GRAPH NORMALISATION

E.g. lemmatisation



PREPROCESSING: INDIRECTED GRAPH NORMALISATION

E.g. synonyms, string similarity



HACKING

HACKING

Tasks:

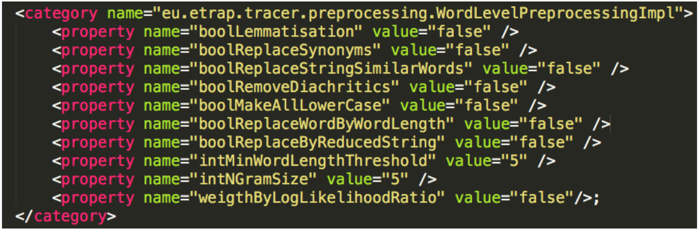
- Run on your texts ...
    1. ... without preprocessing
    2. ... 1) + lemmatisation
    3. ... 2) + synonym replacement

HACKING

Questions:

- Compare the input file with the `*.prep` file for all preprocessing techniques. Which methods seem to work best for you? Which make no sense for the dataset?
- Compare all `*.meta` files containing some numbers! How many words have changed and through which method?
- (optional and advanced) What is the number of word types for each preprocessing technique (can be derived from the first column of `*.prep.inv`).

PREPROCESSING 1) WITHOUT PREPROCESSING

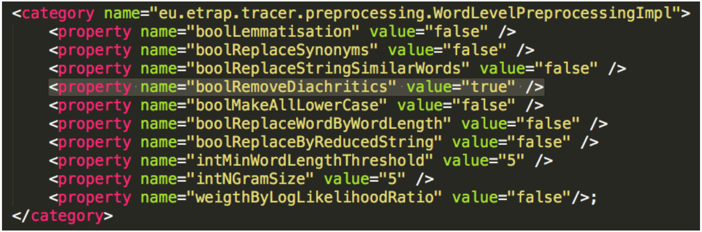


```xml
<category name="eu.etrap.tracer.preprocessing.WordLevelPreprocessingImpl">
    <property name="boolLemmatisation" value="false" />
    <property name="boolReplaceSynonyms" value="false" />
    <property name="boolReplaceStringSimilarWords" value="false" />
    <property name="boolRemoveDiachritics" value="false" />
    <property name="boolMakeAllLowerCase" value="false" />
    <property name="boolReplaceWordByWordLength" value="false" />
    <property name="boolReplaceByReducedString" value="false" />
    <property name="intMinWordLengthThreshold" value="5" />
    <property name="intNGramSize" value="5" />
    <property name="weigthByLogLikelihoodRatio" value="false"/>;
</category>
```

Hint:

- The configuration file can be found in: `$TRACER_HOME/conf/tracer_conf.xml`
- All values show `false`.

PREPROCESSING 2) REMOVING DIACHRITICS

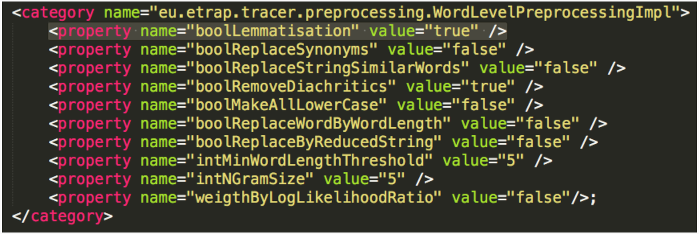


```
<category name="eu.etrap.tracer.preprocessing.WordLevelPreprocessingImpl">
    <property name="boolLemmatisation" value="false" />
    <property name="boolReplaceSynonyms" value="false" />
    <property name="boolReplaceStringSimilarWords" value="false" />
    <property name="boolRemoveDiachritics" value="true" />
    <property name="boolMakeAllLowerCase" value="false" />
    <property name="boolReplaceWordByWordLength" value="false" />
    <property name="boolReplaceByReducedString" value="false" />
    <property name="intMinWordLengthThreshold" value="5" />
    <property name="intNGramSize" value="5" />
    <property name="weigthByLogLikelihoodRatio" value="false"/>;
</category>
```

Hint:

- `boolRemoveDiachritics` is switched on by value `true`.

PREPROCESSING 4) LEMMATISING TEXT

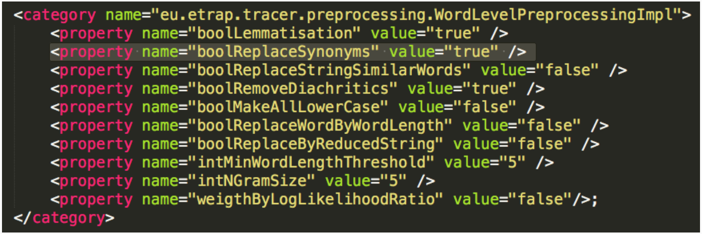


```
<category name="eu.etrap.tracer.preprocessing.WordLevelPreprocessingImpl">
    <property name="boolLemmatisation" value="true" />
    <property name="boolReplaceSynonyms" value="false" />
    <property name="boolReplaceStringSimilarWords" value="false" />
    <property name="boolRemoveDiachritics" value="true" />
    <property name="boolMakeAllLowerCase" value="false" />
    <property name="boolReplaceWordByWordLength" value="false" />
    <property name="boolReplaceByReducedString" value="false" />
    <property name="intMinWordLengthThreshold" value="5" />
    <property name="intNGramSize" value="5" />
    <property name="weigthByLogLikelihoodRatio" value="false"/>;
</category>
```

Hint:

- `boolLemmatisation` is switched on by value `true`.
- **Lemmatisation** can be configured by:
  `<property name="BASEFORM_FILE_NAME" value="data/corpora/Bible/Bible.lemma" />`

PREPROCESSING 5) SYNONYM HANDLING

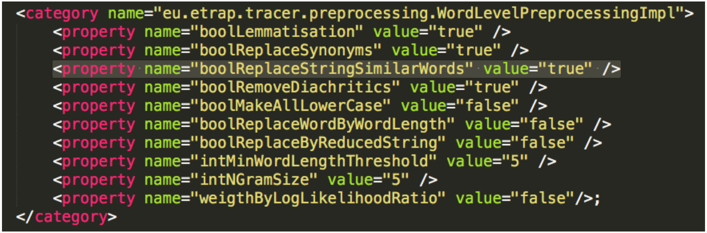


```
<category name="eu.etrap.tracer.preprocessing.WordLevelPreprocessingImpl">
    <property name="boolLemmatisation" value="true" />
    <property name="boolReplaceSynonyms" value="true" />
    <property name="boolReplaceStringSimilarWords" value="false" />
    <property name="boolRemoveDiachritics" value="true" />
    <property name="boolMakeAllLowerCase" value="false" />
    <property name="boolReplaceWordByWordLength" value="false" />
    <property name="boolReplaceByReducedString" value="false" />
    <property name="intMinWordLengthThreshold" value="5" />
    <property name="intNGramSize" value="5" />
    <property name="weigthByLogLikelihoodRatio" value="false"/>;
</category>
```

Hint:

- `boolReplaceSynonyms` is switched on by value `true`.
- **Synonyms** can be configured by:
  `<property name="SYNONYMS_FILE_NAME"`
  `value="data/corpora/Bible/Bible.syns" />`

PREPROCESSING 6) STRING SIMILARITY FOR NORMALISING VARIANTS



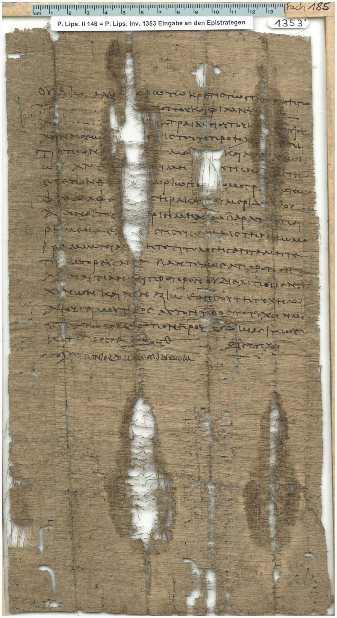
```
<category name="eu.etrap.tracer.preprocessing.WordLevelPreprocessingImpl">
    <property name="boolLemmatisation" value="true" />
    <property name="boolReplaceSynonyms" value="true" />
    <property name="boolReplaceStringSimilarWords" value="true" />
    <property name="boolRemoveDiachritics" value="true" />
    <property name="boolMakeAllLowerCase" value="false" />
    <property name="boolReplaceWordByWordLength" value="false" />
    <property name="boolReplaceByReducedString" value="false" />
    <property name="intMinWordLengthThreshold" value="5" />
    <property name="intNGramSize" value="5" />
    <property name="weigthByLogLikelihoodRatio" value="false"/>;
</category>
```

Hint:

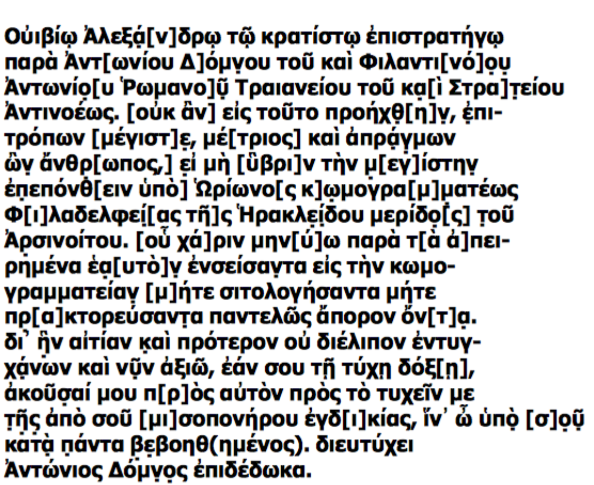
- `boolReplaceStringSimilarWords` is switched on by value `true`.
- Thresholds:

  `<property name="SYNONYMS_FILE_NAME" value="data/corpora/Bible/Bible.syns" />`

OPEN ISSUE: FRAGMENTARY WORDS

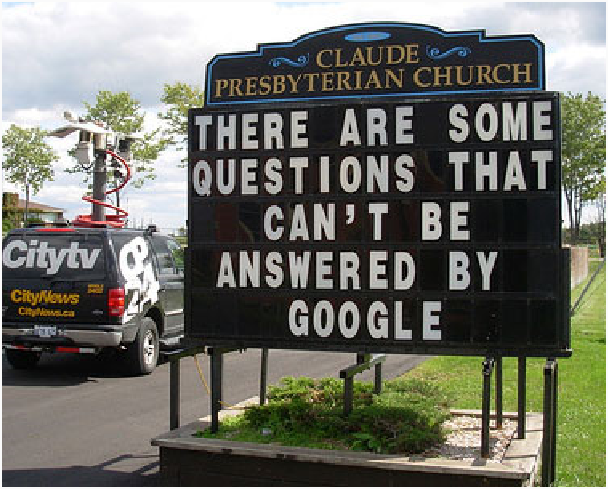

Fach 185
P. Lips. II 146 = P. Lips. Inv. 1353 Eingabe an den Epistrategen
1353

OPEN ISSUE: FRAGMENTARY WORDS - DEALING WITH GAPS AND LEIDEN CONVENTIONS



Οὐιβίῳ Ἀλεξά̣[ν]δρῳ τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμν̣ου τοῦ καὶ Φιλαντι[νό]ο̣υ̣
Ἀντωνίο̣[υ Ῥωμανο]ῦ̣ Τραιανείου τοῦ κα̣[ὶ Στρα]τ̣είου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ̣[η]ν̣, ἐ̣πι-
τρόπων [μέγιστ]ε̣, μέ[τριος] καὶ ἀπρά̣γμων
ὢν̣ ἄνθρ[ωπος,] ε̣ἰ μὴ [ὕβρι]ν τὴν μ̣[εγ]ίστην̣
ἐ̣πεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ω̣μογρα[μ]ματέως
Φ[ι]λαδελφεί̣[α̣ς τῆ]ς Ἡρακλε̣ί̣δου μερίδο̣[ς] τ̣οῦ
Ἀ̣ρσινοίτου. [οὗ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρ̣ημένα ἑα̣[υτὸ]ν̣ ἐνσείσαν̣τα εἰς τὴν κωμο-
γραμματείαν̣ [μ]ήτε σιτολογήσαντα μήτε
πρ[α]κτορεύσαντ̣α παντελῶς ἄπορον ὄν[τ]α̣.
δι᾽ ἣν αἰτίαν κ̣αὶ πρότερον οὐ διέλιπον ἐντυγ-
χά̣νων καὶ νῦ̣ν ἀξιῶ, ἐάν σου τῆ̣ τύχη̣ δόξ[η̣],
ἀκοῦσ̣αί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με
τ̣ῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας, ἵν᾽ ὦ ὑπὸ̣ [σ]ο̣ῦ̣
κατὰ̣ π̣άντα βε̣βοηθ(ημένος). διευτύχει
Ἀντώνιος Δόμν̣ο̣ς ἐπιδέδωκα.

GAP BETWEEN KNOWLEDGE AND EXPERIENCE


CLAUDE PRESBYTERIAN CHURCH
THERE ARE SOME QUESTIONS THAT CAN'T BE ANSWERED BY GOOGLE
Citytv
CityNews
CityNews.ca

CONCLUSION AND REVISION

CHECK

Statement:

- ”My lemmatisation tool <XYZ> is able to compute the base forms of 80% of all tokens in a corpus.”

Good or bad?

CHECK

Fact file:

- Language variants
- Different writing styles
- (Some) dialects
- Diachritics
- OCR errors

Question: What's the difference for you?

CHECK

Fact file:

- Language variants
- Different writing styles
- (Some) dialects
- Diachritics
- OCR errors

Question: What do you think is the difference for the computer?

IMPORTANCE OF PREPROCESSING

- Cleaning and harmonising the data.
- When working with a new corpus -not only the language but also the same language in different epochs or geographical regions- cleaning/harmonising the data can take up to 70% of the overall time.

Preprocessing mantra:
Garbage in, garbage out

Finito!

CONTACT

Team

Marco Büchler, Greta Franzini and Emily Franzini.

Visit us

🌐 `http://www.etrap.eu`

✉ `contact@etrap.eu`

TRACER
TEXT REUSE DETECTION MACHINE

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

SPONSORED BY THE
Federal Ministry
of Education
and Research

LICENCE