Natural Language Processing
SoSe 2015

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Exercise 1: Language Modelling

Dr. Mariana Neves

April 20th, 2015

Genia corpus

- Genia corpus (Treebank)
  - Around 2,000 abstracts in XML format
  - Biomedical domain
  - Sentence split, tokenized, POS tagged

- Available at:
  - http://www.nactem.ac.uk/genia/genia-corpus/treebank

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE Annotation>
<?xml-stylesheet type="text/css" href="GENIA_tree_bracket.css"?>
<Annotation>
<PubmedArticleSet>
<PubmedArticle>
<MedlineCitation>
<PMID>1309833</PMID>
<Article>
<ArticleTitle><sentence id="S1"><cons cat="NP"><cons cat="NP"><cons cat="NP"><tok cat="NN">Cortisol</tok> <tok cat="
<Abstract>
<AbstractText><sentence id="S2"><cons cat="S"><cons cat="NP" id="i33" role="SBJ"><tok cat="JJ">Primary</tok> <tok ca
<sentence id="S3"><cons cat="S"><cons cat="NP" id="i34" role="SBJ"><tok cat="PRPP">Its</tok> <tok cat="NN">occurrence
<sentence id="S4"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><tok cat="DT">the</tok> <tok ca
<sentence id="S5"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">The</tok> <tok cat="JJ">first</tok> <tok cat
<sentence id="S6"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="NN">Hydrochlorothiazide</tok> <tok cat="NN">therap
<sentence id="S7"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">The</tok> <tok cat="JJ">second</tok> <tok ca
<sentence id="S8"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">All</tok> <tok cat="CD">four</tok> <tok cat="
<sentence id="S9"><cons cat="S"><cons cat="NP" role="SBJ"><cons cat="NP"><tok cat="JJ">Low</tok> <tok cat="NN">dose</
<sentence id="S10"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">All</tok> <tok cat="NNS">patients</tok></con
<sentence id="S11"><cons cat="S"><cons cat="NP" role="SBJ"><cons cat="NP"><tok cat="DT">The</tok> <tok cat="JJ">diurn
<sentence id="S12"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="EX">There</tok></cons> <cons cat="VP"><tok cat=
<sentence id="S13"><cons cat="S"><cons cat="NP" id="i36" role="SBJ"><tok cat="VBG">Circulating</tok> <tok cat="JJ">ad
<sentence id="S14"><cons cat="S"><cons cat="NP" id="i37" role="SBJ"><tok cat="NN">Glucocorticoid</tok> <tok cat="NNS"
<sentence id="S15"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><tok cat="DT">the</tok> <tok c
<sentence id="S16"><cons cat="S"><cons cat="NP" id="i38" role="SBJ"><cons cat="NP"><tok cat="DT">A</tok> <tok cat="J
<sentence id="S17"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><tok cat="DT">the</tok> <tok c
<sentence id="S18"><cons cat="S"><cons cat="PP"><tok cat="IN">As</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">a
<sentence id="S19"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">t
<sentence id="S20"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="JJ">Partial</tok> <tok cat="NN">cortisol</tok> <
<sentence id="S21"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">t
<sentence id="S22"><cons cat="S"><cons cat="NP" role="SBJ"><cons cat="NP"><tok cat="NN">Therapy</tok></cons> <cons c
</Abstract>
</Article>
</MedlineCitation>
</PubmedArticle>
</PubmedArticleSet>
</Annotation>
```

Genia corpus

```
<sentence id="S9">
<cons cat="S">
    <cons cat="NP" role="SBJ">
        <cons cat="NP">
            <tok cat="JJ">Low</tok>
            <tok cat="NN">dose</tok>
            <tok cat="NN">dexamethasone</tok>
            <tok cat="NN">therapy</tok>
        </cons>
        <cons cat="PRN">
            <tok cat="LRB">(</tok>
            <cons cat="NP">
                <cons cat="QP">
                    <tok cat="CD">1-1.5</tok>
                </cons>
                <tok cat="NN">mg/day</tok>
            </cons>
            <tok cat="RRB">)</tok>
        </cons>
    </cons>
<cons cat="VP">
    <tok cat="VBD">was</tok> ...
</cons>
</sentence>
```

word

Task: Language modelling

- Calculate the perplexity of the corpus
- Split in training (90%) and test (10%) sets
  - Randomly and per document
  - Build a language model using the training set (Maximum Likelihood, bigram)
  - Use Laplace (add-one) smoothing for the zero probabilities
  - Estimate probabilities on the test set

Exercise 1

- Deadline on May 11th, Upload to HPI owncloud
  - README file with comments, instructions and results (perplexity/precision)
  - Zipped source code

- Late submissions will lose 0.5/20 point in the exam