Data visualization – Text

The first step

- Text has to be turned into data
- Quantitative coding
    - Human coding
    - „automatic“ (machine processed) coding
        - Almost always inclde some decision of human
        - Frequency analysis (Wordclouds)
        - Sentiment analysis - Is the text positive or negative?
        - Topic models - what topics are in the text(s)
        - Analysis of readability (how complex the text is?)
        - Analysis of „positions“

- Allocating recording units to substantial categories
    - Classifying each coded unit of text from the sample according to the category scheme
        - Coding scheme
            - Before coding (deductive)
            - During coding (inductive)
        - Unit: page, paragraph, sentence, quasisentence

Quasi-sentence

- an argument or phrase which is the verbal expression of one idea or issue
- One sentence may include more ideas, one idea may be divided into more sentences.

- I am going to buy bread, milk and apple.

- I am going to buy bread.
- I am going to buy milk.
- I am going to buy apple.

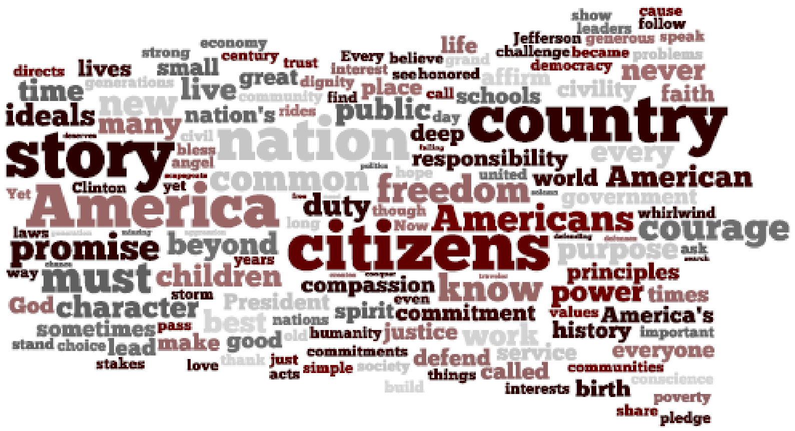
Text elements

- Paragraphs
  - Sentences
    - N-grams
      - Words
        - Lemma
        - token

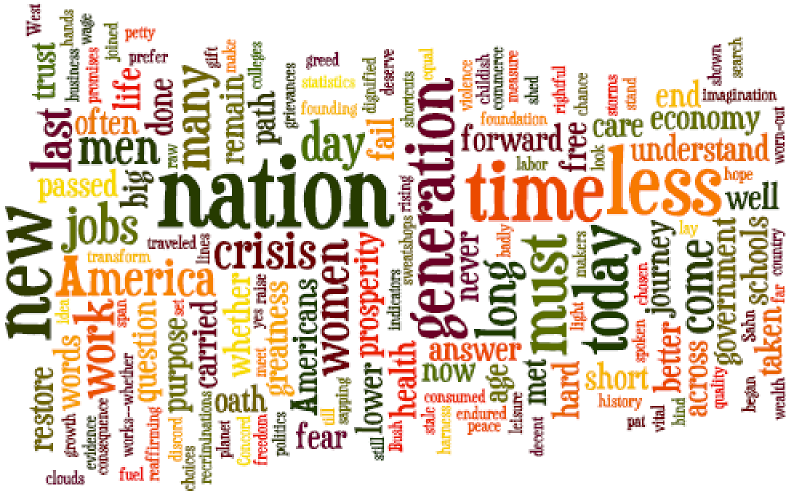
Problems

- Language
- Tenses, adjective, male/female
- Common words (stopwords)
- Reliability of human coding
- Different understaning by different coders
- Noise in automatic coding

Wordcloud



(c) Bush



(d) Obama