Natural Language Processing
(11. přednáška)

Intro (coursera.org)

```
https:
//class.coursera.org/nlp/lecture/preview_view/124
```

Aplikace: detekce plagiátů

Dva možné přístupy:

Aplikace: detekce plagiátů

Dva možné přístupy:

- analýza dokumentu

Aplikace: detekce plagiátů

Dva možné přístupy:

- analýza dokumentu
  - stylometrie

Aplikace: detekce plagiátů

Dva možné přístupy:

- analýza dokumentu
  - stylometrie(autorský invariant

Aplikace: detekce plagiátů

Dva možné přístupy:

- analýza dokumentu
  - stylometrie(autorský invariant, genetické algoritmy)

Aplikace: detekce plagiátů

Dva možné přístupy:

- analýza dokumentu
  - stylometrie(autorský invariant, genetické algoritmy)
  - neuronové sítě
  - porovnávání dokumentu s databází jiných dokumentů

Porovnávání dokumentů: Špatný přístup

Chceme zjistit, zda mají dva dokumenty společné části.

Porovnávání dokumentů: Špatný přístup

Chceme zjistit, zda mají dva dokumenty společné části.

```
def longest_match(A,B):
  ret = ''
  i=0
  while A[i] == B[i]:
    ret += A[i]
    i += 1
  return ret

def common( A, B ):
  longest = ''
  for a_pos in len(A):
    for b_pos in len(B):
      m = longest_match(A[a_pos:], B[b_pos:])
      if len(m) > longest:
        longest = m
```

Porovnávání dokumentů: Dynamické programování

Pro každou dvojici prefixů spočti nejdelší společný suffix. Nejdelší ze všech těchto suffixů je nejdelší společný string dvou dokumentů

Porovnávání dokumentů: Dynamické programování

Pro každou dvojici prefixů spočti nejdelší společný suffix. Nejdelší ze všech těchto suffixů je nejdelší společný string dvou dokumentů

Složitost $O(nm)$

Porovnávání dokumentů: Suffix trees

- stromová struktura odpovídající řetězci

Porovnávání dokumentů: Suffix trees

- stromová struktura odpovídající řetězci
- hrany stromu jsou ohodnoceny podřetězci

Porovnávání dokumentů: Suffix trees

- stromová struktura odpovídající řetězci
- hrany stromu jsou ohodnoceny podřetězci
- každý suffix odpovídá právě jedné cestě od kořenu k listu

Porovnávání dokumentů: Suffix trees

- stromová struktura odpovídající řetězci
- hrany stromu jsou ohodnoceny podřetězci
- každý suffix odpovídá právě jedné cestě od kořenu k listu
- umožňují rychle provádět se striny některé operace

Porovnávání dokumentů: Suffix trees

- stromová struktura odpovídající řetězci
- hrany stromu jsou ohodnoceny podřetězci
- každý suffix odpovídá právě jedné cestě od kořenu k listu
- umožňují rychle provádět se striny některé operace
- zobecnění umožňuje najít LCS v čase $O(n + m)$!

Porovnávání dokumentů: “otisky”

Ke každému dokumentu si uložím množinu všech n-gramů

-

Winnowing

Nepodstatné změny