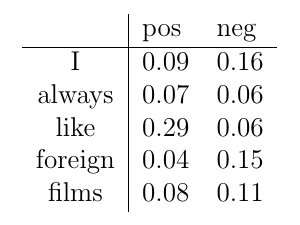
Zadanie 0.:

Dlaczego jednym z częstych kroków preprocessingu jest usuniecie wielkich liter? Podaj przykład zadania NLP, dla którego usuniecie wielkich liter jest niekorzystne.

Usuwamy wielkie litery ponieważ algortym jest CaseSensitive – wrażliwy na wielkość liter. Zliczanie częstotliwości występowania słów będzie nieprawdziwe jeśli zachowamy wielkie litery (np. na początku zdania). Jest to niekorzystne ze względu na nazwy własne – wtedy algorytmy mają utrudnione zadanie w rozpoznawaniu ich.

Zadanie 1.:

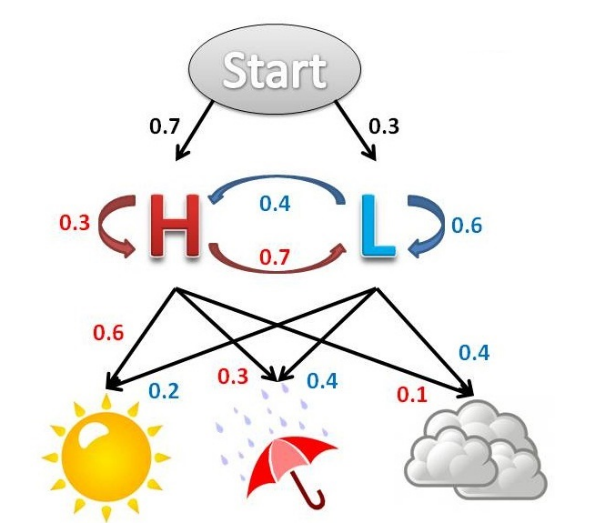
Jaka jest najbardziej prawdopodobna klasa dla zdania `**`I always like foreign films**'' wg naiwnego klasyfikatora Bayesowskiego? Podaj odpowiednie prawdopodobieństwa i typowaną klasę.

0,09\*0,07\*0,29\*0,04\*0,08 = 0,0000058464

0,16\*0,06\*0,06\*0,15\*0,11 = 0,000009504

**P(neg) > P(pos)**

Zadanie 2.:

(1p) Zapisz wszystkie elementy konieczne do zdefiniowania powyższego HMM.

-Macierz początkowa i przejść

|  |  |  |
| --- | --- | --- |
|  | H | L |
| <s> | 0,7 | 0,3 |
| H | 0,3 | 0,7 |
| L | 0,6 | 0,4 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Słońce | Deszcz | Chmury |
| H | 0,6 | 0,3 | 0,1 |
| L | 0,2 | 0,4 | 0,4 |

-Macierz emisji

(3p) Jakie prawdopodobieństwo ma zdarzenie **Rainy-Sunny-Cloudy**?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Rainy | Sunny | Cloudy |
| H | 0,21 | 0,0378 0,0441 | 0,001323 0,003087 |
| L | 0,12 | 0,0144 0,0096 | 0,002304 0,003456 |

(3p) Jakie najbardziej prawdopodobny ciąg stanów ukrytych wygenerował powyższą obserwację?

H -> H -> L

Zadanie 3.:

Zbiór dokumentów:

doc1: “Ave Caesar, morituri te salutant”

doc2: “Ave, Imperator, morituri te salutant”

doc3: “Ave Caesar nos morituri te salutamus”.

Dla każdego z tych dokumentów wypisz jego reprezentacje Bag of Words oraz Set of Words.

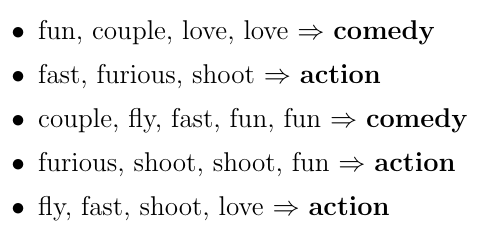
BoW(doc1): “Ave”:1, “Caesar”:1, “,”:1, “morituri”:1, “te”:1, “salutant”:1  
SoW(doc1): “Ave”, “Caesar”, “,”, “morituri”, “te”, “salutant”

BoW(doc2): “Ave”:1, ”,”:2, ”Imperator”:1, “morituri”:1, “te”:1, “salutant”:1  
SoW(doc2): “Ave”, ”,”, ”Imperator”, “morituri”, “te”, “salutant”

BoW(doc3): “Ave”:1, “Caesar”:1, “nos”:1, “morituri”:1, “te”:1, “salutamus”:1  
SoW(doc3): “Ave”, “Caesar”, “nos”, “morituri”, “te”, “salutamus”

Zadanie 4.:

Jaka jest najbardziej prawdopodobna klasa dla dokumentu o cechach

``fast, couple, shoot, fly''

wg naiwnego klasyfikatora Bayesowskiego? Załóż, że klasyfikator wykorzystuje wygładzanie ze stałą \alpha=1. Przedstaw pełne rozwiązanie oraz podaj odpowiednie prawdopodobieństwa i typowaną klasę.

P(comedy) = 2/5 Ilość słów w comedy = 5

P(action) = 3/5 Ilość słów w action = 7

Alpha = 1

P(comedy|”fast,couple,shot,fly”) = 2/5 \* 2/12 \* 3/12 \* 1/12 \* 2/12 =   
 = 24/103680 = 2,315\*10E-4

P(action|”fast,couple,shot,fly”) = 3/5 \* 3/13 \* 1/13 \* 5/13 \* 2/13 =   
 = 90/142805 = 6,302\*10E-4

P(action|”[…]”) > P(comedy|”[…]”)

Klasa typowana to **“action”.**

Zadanie 5.:

Podaj 5 przykładów stop wordsów dla języka polskiego.

Stopwords(pl): “co”, “do”, “mu”, "mi”, "im”

Zadanie 6.:

Test diagnostyczny na chorobę X ma prawdopodobieństwo 0.95 dania wyniku pozytywnego po zbadaniu osoby chorej na chorobę X, oraz 0.10 dania wyniku pozytywnego po zbadaniu osoby zdrowej. Estymuje się, ze 0.5% populacji cierpi na chorobę X. Załóżmy, ze chcemy się zbadać, czy jesteśmy chorzy nie posiadając żadnych istotnych informacji wpływających na prawdopodobieństwo bycia chorym (poza faktem, ze należymy do populacji). Oblicz prawdopodobieństwa, ze:

|  |  |  |
| --- | --- | --- |
|  | Wynik pozytywny | Wynik negatywny |
| Osoba chora | 0,95 | 0,05 |
| Osoba zdrowa | 0,10 | 0,90 |

1. wynik testu będzie pozytywny

0,05\*0,95 + 0,05\*0,1 = 0,0475 + 0,005 = 0,0525

2. pod warunkiem pozytywnego wyniku, jesteśmy chorzy na chorobę X

0,0525\*0,95 = 0,049875

3. pod warunkiem negatywnego wyniku, nie jesteśmy chorzy na chorobę X

0,9475\*0,9 = 0,85275

4. zostaniemy źle sklasyfikowani

0,05\*0,05 + 0,05\*0,90 = 0,0025 + 0,045 = 0,0475