# Tema 3 IA 2018

Sumarizare de texte

Vlad Bogolin

## Scopul temei

Scopul temei este acela de familiarizare cu tehnici de baza din prelucrarea limbajului natural. Prelucrarea limbajului natural este una dintre principalele direcţii de cercetare în momentul actual. Sumarizarea textelor consta în crearea unui text scurt şi concis care sa reprezinte rezumatul unui document mai lung. Asadar, sumarizarea automata este o problema din ce in ce mai importanta datorita numarului mare de texte (de exemplu, stiri) ce se gasesc online. Metodele de sumarizarea automata ajuta la descoperirea informatiei relevante.

## Enunt

Dandu-se un articol de stiri (titlu + text), sa se produca un rezumat al acestuia care sa fie cat mai relevant. Algoritmul va avea urmatorii pasi:

1. Preprocesarea articolului - eliminarea informatiei nerelevante (cuvinte de stop, numere, etc). Pentru acest pas se gasesc online liste ce contin cuvintele de stop.
2. Definire vocabular - pentru a limita numarul de cuvinte folositi lematizare/stemming si puteti ignora cuvintele ce apar de putine ori. De asemenea, puteti lua in calcul doar substantivele (vezi cerinta 3).
3. Calculare Term Frequency (TF) pentru fiecare cuvant din vocabular
4. Calculare TF-IDF (Term Frequency - Inverse Document Frequency) pentru fiecare cuvant din vocabular
5. Calculare scor pentru fiecare propozitie din articolul de interes
6. Creare rezumat folosind propozitiile cu cel mai mare scor

## Cerinta 1 [2p] - Curatare text

Eliminare cuvinte de stop, semne de punctuatie, eventuale spatii suplimentare. Aceasta cerinta se va implementa intr-o functie separata care primeste ca parametru textul unui articol si intoarce textul preprocesat.

# Cerinta 2 [3p] - Calculare TF-IDF

Calculare TF-IDF pe baza textelor din setul de date de antrenare pentru fiecare cuvant din vocabular. Aveti grija sa preprocesati textul inainte (lematizare/stemming - pentru lematizare/stemming puteti folosi orice resursa, etc).

# Cerinta 3 [3p] - Calculare scor per propozitie

Folosind valorile TF-IDF calculate pana acum, se va calcula un scor pentru fiecare propozitie. Pentru impartirea in propozitii puteti folosi nltk sau orice altceva. Scorul pentru o propozitie va fi dat de suma TF-IDF-ului pentru fiecare cuvant. Totusi, pentru a avea un scor cat mai bun se vor parcurge urmatorii pasi:

1. Se va calcula scorul doar pentru substantive. Pentru a identifica partile de vorbire puteti folosi orice resurse (nltk, etc.). Adica scorul propozitiei va fi afectat doar de scorul substantivelor continute. Normalizati scorul propozitiei.
2. Calculare similaritate cu titlul. Mai exact, se va adauga o valoare aditionala daca propozitia are cuvinte ce apar in titlu. Aceasta valoare este egala cu numarul de cuvinte din propozitie care se regasesc in titlu impartit la numarul total de cuvinte din titlu. La final, aceasta valoare se va pondera cu o constanta care ii defineste importanta (de exemplu 0.1) si se va aduna la scorul calculat la pasul 1.
3. La final se va pondera fiecare propozitie cu o pondere (intre 0 si 1) corespunzatoare cu pozitia ei in text. De exemplu, pentru un text cu 10 propozitii, ponderea pentru cea de-a noua va fi 0.9. Acest lucru se intampla deoarece se tinde ca propozitiile cele mai importante sa fie spre finalul articolului (de exemplu concluzia). Definiti propria metrica de ponderare a propozitiei in functie de locatia in document.
4. Se vor returna primele 3 propozitii cu cel mai mare scor.

# Cerinta 4 [2p] - Evaluare rezultate

Pentru evaluare se va folosi ground truth-ul pentru fiecare propozitie din dataset. Avand in vedere ca sumarizarea nu foloseste informatia de ground truth, evaluare si invatarea TF-IDF-ului se pot face pe acelasi dataset.

Pentru evaluare se vor folosi doua metrici: $BLEU@n$ si $ROUGE@n$. In general metrica BLEU masoara precizia, iar ROUGE recall-ul. Ele sunt definite dupa cum urmeaza:

- $BLEU@n = \frac{nr\ n-grame\ comune\ intre\ text\ si\ ground\ truth}{nr\ n-grame\ in\ text}$
- $ROUGE@n = \frac{nr\ n-grame\ comune\ intre\ text\ si\ ground\ truth}{nr\ n-grame\ in\ ground\ truth}$

Realizati grafice sau tabele pentru a scoate in evidenta cum influenteaza scorul fiecare pas din algoritm (cel putin urmatoarele: cum se comporta algoritmul fara fiecare din cei 3 pasi, cum influenteaza ponderea similaritatii cu titlul performanta, cum influenteaza politica de ponderare a propozitiei in raport cu locatia ei in text performanta). In comparatie se vor folosi folosi metricile pentru unigrame, bigrame si 4-grame, adica (BLEU@1, BLEU@2, BLEU@4 si ROUGE@1, ROUGE@2 si ROUGE@4). Pentru a calcula scorul pe tot datasetul se va face media scorurilor pentru fiecare text in parte.

# Bonus [max 2p]

Aduceti imbunatatiri proprii algoritmului. Punctajul pentru bonus se va acorda in functie de complexitatea imbunatatirilor aduse cat si de performanta acestora.

# Descriere set de date

Setul de date pentru evaluare este in format csv si contine urmatoarele coloane:

- headlines - titlul
- text - sumarizarea textului (ground truth, se va folosi doar pentru evaluare)
- ctext - textul complet

Datasetul se gaseste aici:
https://drive.google.com/open?id=1jNqEdgHYcrX9aa41guTvJtUxyB5-fnuy

# Alte precizari

In afara de cazurile specificate in enunt, nu se pot folosi resurse externe pentru a realiza tema.