**Mecanism personalizat de căutare bazat pe conținut pentru videoclipuri educaționale**

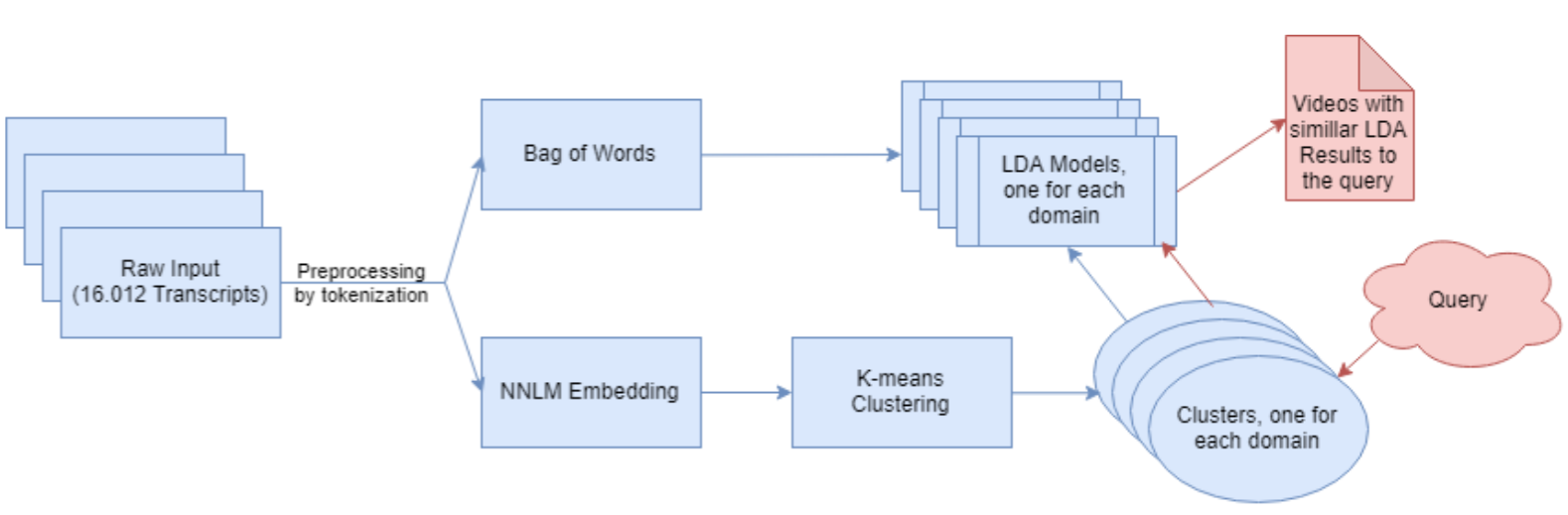
Turcu Gabriel

Facultatea de Automatică, Calculatoare si Electronică

**Rezumat**: Găsirea videoclipurilor potrivite dintr-un repertoriu online de resurse video educaționale este foarte importantă pentru dezvoltarea studenților. Considerând că avem la dispoziție transcrierile pentru toate videoclipurile, problema se reduce la obținerea unei liste sortate cu cele mai bune rezultate pentru o anumita interogare. Lucrarea prezinta un mecanism personalizat de căutare pentru cele 16.012 de transcrieri ale videoclipurilor de pe website-ul <https://media.upv.es/> al Universității Politehnice din Valencia. O problema a acestui mecanism personalizat de căutare provine de la faptul că transcrierile sunt in limba spaniolă. Soluția propusă aplică un algoritm de word embedding folosind Neural-Net Language Models pe toate transcrierile, face clustering ca să separe videoclipurile in funcție de domeniul din care fac parte si apoi construiește un model Latent Dirichlet Allocation (LDA) pentru fiecare cluster. La urma, putem să procesăm o interogare noua si să găsim videoclipurile cele mai relevante cu căutarea noastră. Transcrierile celor mai relevante videoclipuri o să facă parte din același cluster ca si interogarea si o să aibă scorurile LDA cele mai apropriate de scorurile interogării.

**Introducere**: In ultimii ani, creșterea gradului de utilizare al platformelor de tip Massive Online Open Courses (MOOCs), cât si cele de tip Technology Enhanced Learning (TEL) au evidențiat nevoia si importanta unui sistem eficient si precis de căutare pentru videoclipuri educaționale. Creșterea gradului de utilizare al platformelor menționate a adus cu sine si creșterea cantității de informație prin care trebuie căutat, informație care de cele mai multe ori nu este marcată corespunzător (titlu derutant, cuvinte cheie lipsă, etichete lipsă, etc.), lucru care creste nevoia pentru un mecanism de căutare bazat pe elemente de conținut deja existente si care nu mai au nevoie de preprocesare manuala (adăugarea manuala de cuvinte cheie, etichete, clasificare pe categorii, etc.).

**Descrierea soluției**: Soluția propusă preprocesează transcrierile folosind procesul de tokenizare, după care aplica un algoritm de word embedding folosind Neural-Net Language Models pe toate transcrierile.



***Fig. 1. Arhitectura sistemului***

Pe rezultatele de la word embedding se aplică K-Means clustering ca să separe videoclipurile in 4 clustere in funcție de domeniul din care fac parte (Arta si Umanitate, Inginerie si Arhitectura, Științe (Biologice) sau Social si Legal ) si apoi construiește un model Latent Dirichlet Allocation (LDA) pentru fiecare din cele 4 clustere. La urma, putem să procesăm o interogare noua ca să aflam din ce cluster face parte si apoi si să găsim videoclipurile ale căror transcrieri o să facă parte din același cluster si ale căror scoruri LDA sunt cele mai apropriate de scorurile interogării.

**Rezultate experimentale**: Pentru realizarea acestei soluții, am folosit Python ca limbaj de programare si următoarele librarii pentru implementare: sklearn pentru KMeans clustering, gensim pentru LDA si tensorflow pentru partea de NNLM word embeddings. In urma interogării, putem să vedem care sunt ID-urile videoclipurilor care sunt cele mai apropriate ca scor adică cele care au conținutul cel mai relevant la interogarea noastră.

|  |  |  |  |
| --- | --- | --- | --- |
| Interogare | Cluster  Asignat | Scoruri LDA  (ID Topic si Scor) | Lista transcrierilor sortate in funcție de relevanta  (ID Transcriere si Scor) |
| Ciencias de la Computacion | 2 | (0, 0.4422) (1, 0.0530)  (2, 0.0443) (6, 0.0335)  (7, 0.4258) ... | (41107,0.1097) (765,0.1120) (18460,0.1124) (3895,0.1141) (1236,0.1146) ... |
| permanente para la proteccion de los animales en cria instituido | 2 | (6,0.01429) (7,0.871) (8,0.01428) (9,0.01428) ... | (44794,0.0012) (27277,0.0017) (44466,0.032) (26721,0.0428) (41793,0.0428)... |

**Concluzii**: In acest proiect am realizat o procedura personalizata de căutare bazata pe conținut pentru cele 16.012 transcrieri folosind word embeddings, K-Means clustering si LDA. Ulterioare îmbunătățiri ale proiectului ar putea însemna obținerea datelor privind adevărul despre clasificarea transcrierilor sau despre numărul de domenii sau despre numărul de teme pe care ar trebui să le detecteze algoritmul LDA.

Aceasta aplicație a fost dezvoltata in cadrul unei mobilități Erasmus+ de doua luni la Universitatea Politehnica din Valencia. Pe lângă aplicația propriu zisă a mai fost realizata si o lucrare științifica care a fost trimisă la conferința internaționala *The International Conference on Hybrid Artificial Intelligent Systems (HAIS)*, <http://2019.haisconference.eu/>.

**Referințe**:

* Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
* . Iyer, R.R., Parekh, S., Mohandoss, V., Ramsurat, A., Raj, B., Singh, R.: Content-based video indexing and retrieval using corr-lda. arXiv preprint arXiv:1602.08581 (2016)
* Ngo, C.W., Jiang, Y.G., Wei, X.Y., Wang, F., Zhao, W., Tan, H.K., Wu, X.: Experimenting vireo-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search. In: TRECVID (2007)
* Toby Crayston: The Natural Language Processing API. https://www.textrazor.com/technology (Apr. 2019)