AKADEMIA E FORCAVE
TË ARMATOSURA

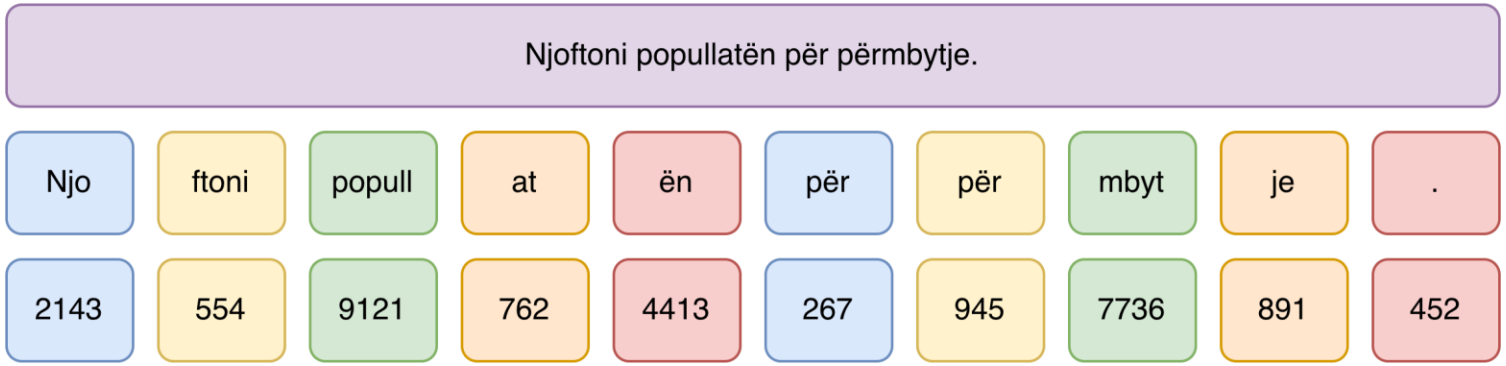
Hyrje në Modele të Mëdha Gjuhësore

Tokenizimi, paraqitjet vektoriale dhe kuptimi semantik

Dr. Fiorela Ciroku

Si e përfaqësojnë MMGj-të tekstin?

- Tokenizimi është procesi i ndarjes së tekstit në njësi më të vogla të quajtura token.
- Këta token mund të jenë fjalë të plota, nënfjalë (subwords) ose edhe karaktere.
- MMGj-të operojnë mbi token, jo mbi tekst të papërpunuar.
- MMGj-të nuk e shohin tekstin si kuptim apo fjali.



Çfarë është tokenizimi?

- Tokenizimi është procesi i konvertimit te tekstit në një sekuencë me copëza më të vogla të cilat mund të procesohen nga MMGj.
- Përkufizon strukturën e inputit për modelin.
- Ndikon në kosto dhe shpejtësi (më shumë token -> përgjigje më e avashtë)
- Influencon saktësinë semantike, pasi segmente teksti të ndara keq e dobësojnë embedding.

Llojet e tokenizimit

- Në nivel karakteri (character-level) -> Çdo karakter është një token.
- Në nivel fjale (word-level) -> Çdo fjalë është një token.
- Në nivel nënfjale (subword)
- Në nivel morfologjik
- Në nivel rregullash (Rule-Based Heuristic)
- Hibrid

Tokenizimi në nivel karakteri

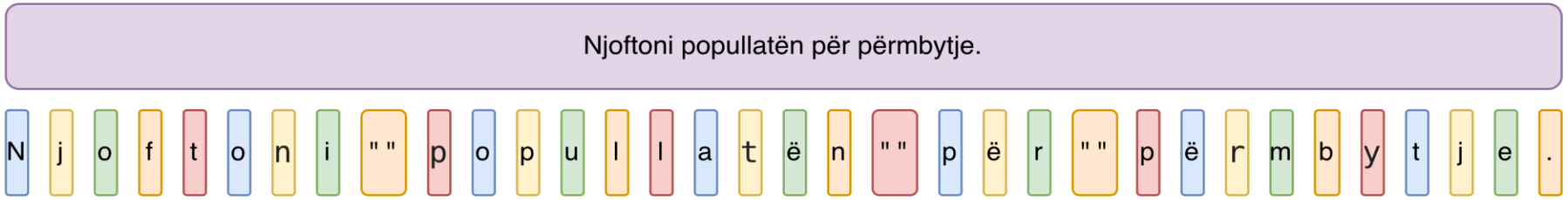
Ndan tekstin ne karaktere te vecanta.

Karakteristika:

- Nuk mbart kuptimin e fjalëve ose nënfjalëve.
- Mund të trajtojë cdo tip teksti.

Kufizime:

- Prodhon sekuenca të gjata që janë ineficente në transformer.
- Humbet strukturën semantike.



Tokenizimi në nivel fjale

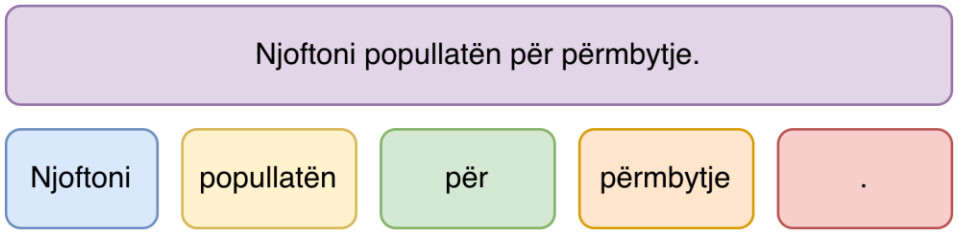
Ndan tekstin në fjalë të plota duke përdorur whitespace ose shenja pikësimi.

Karakteristika:

- Tokenat janë intuitive dhe të lexueshme nga njeriu.
- Madhësia e fjalorit është e madhe.
- Funksionon mirë për gjuhë me kufizime të qarta.

Kufizime:

- Fjalore të mëdha (100k+ tokena)
- Probleme me fjalët e rralla
- Nuk mund të pergjithësojë për injektime të reja.



Tokenizimi në nivel nenfjale

Një kompromis midis nivelit të fjalës dhe nivelit të karakterit.

- Ndan fjalët në njësi domethënëse (morfema, parashtesa, prapashtesa).
- Shmang problemet Out-Of-Vocabulary duke i mbajtur sekuencat më të shkurtra sesa ato me karaktere të veçanta.

Ndahet në:

- Subword Byte Pair Encoding (BPE)
- Byte-level BPE
- WordPiece
- SentencePiece / Unigram LM

Tokenizimi me Subword BPE

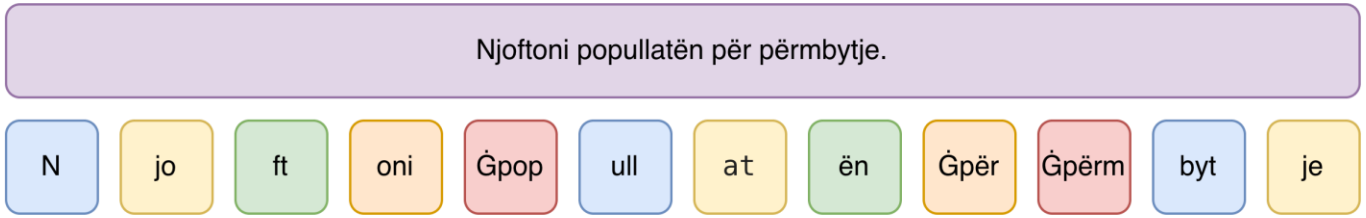
Merr në mënyrë përsëritëse çiftet më të shpeshta të karaktereve ngjitur dhe i bashkon.

Karakteristika:

- determinist dhe i bazuar në frekuencë
- krijon nënfjalë të zakonshme si: **para + laj + më + rim**

***Kufizime*:**

- mund të segmentojë keq fjalët e rralla
- trajtimi i hapësirave bëhet jashtë algoritmit



Tokenizimi me Byte-Level BPE

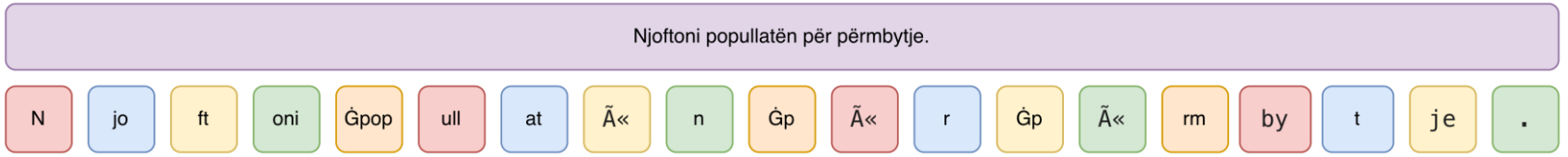
Merr në mënyrë përsëritëse çiftet më të shpeshta të byte-ve ngjitur dhe i bashkon.

Karakteristika:

- Mund të tokenizojë çdo lloj simboli nga çdo lloj gjuhe pa pasur nevojë ta ketë në fjalor.
- Nuk dështon në fjalë të rralla, të reja, emojis, or shkronja të veçanta.

***Kufizime*:**

- Mund të prodhojë disa token për karaktere komplekse, që rezulton në sekuenca tokeni më të gjata.
- Nuk interpretohen kollaj nga njeriu.



Tokenizimi me WordPiece

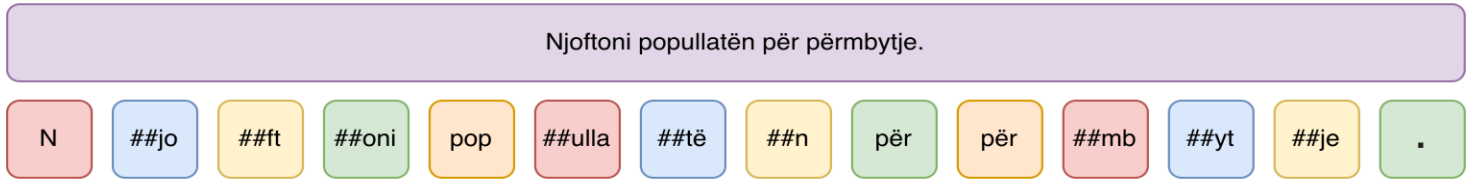
Ndërton fjalorin duke zgjedhur nënfjalë të cilat maksimizojnë mundësinë për rindërtimin e korpusit të trajnimit.

Karakteristika:

- Strategji bashkimi probabilistik
- Shenjues vazhdimi (**##**)

Kufizime:

- Kërkon parapërpunim dhe normalizim të hapësirave.
- Paksa më i ngadaltë për trajnim
- Më pak fleksibël me tekst shumëgjuhësh ose fjalë të rralla.



Tokenizimi me SentencePiece

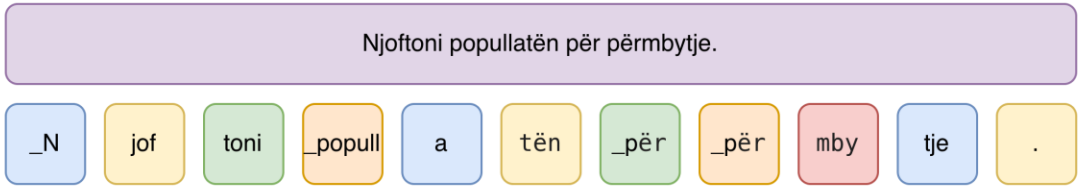
Mëson nënfjalët drejtpërdrejtë nga teksti i papërpunuar (pa nevojë për segmentim të hapësirave), duke përdorur BPE ose model gjuhësor unigram.

Karakteristika:

- trajton hapësirat si karakter (▁)
- Menaxhim i mirë i shumëllojshmërisë së gjuhëve
- Menaxhim i shkëlqyer i Unicode

Kufizime:

- Tokenët mund të jenë më pak intuitivë
- Fjalori varet shumë nga diversiteti i korpusit



Tokenizimi në nivel morfologjik

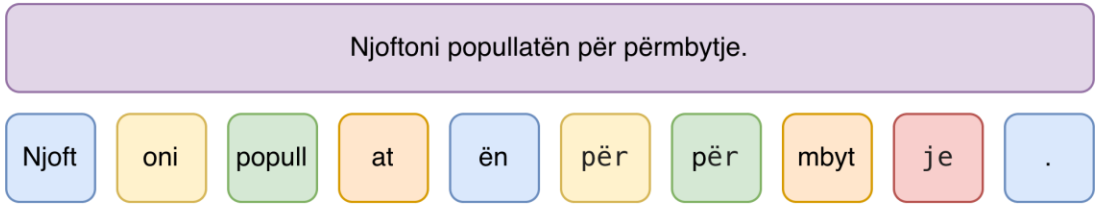
Ndan fjalët në morfema (qeliza më e vogël linguistike) bazuar në morfologjinë e gjuhës që trajton.

Karakteristika:

- Përdor rregulla morfologjik për të ndarë fjalët (parashtesa, prapashtesa, etj).
- Mbart kuptimin semantik dhe informacionin gramatikor.

Kufizime:

- Shumë specifike për gjuhë të caktuara dhe kërkon shume burime informacioni.
- Komplekse për t'u kompjutuar dhe më pak e mundshme të përgjithësohet.



Tokenizimi në nivel rregullash

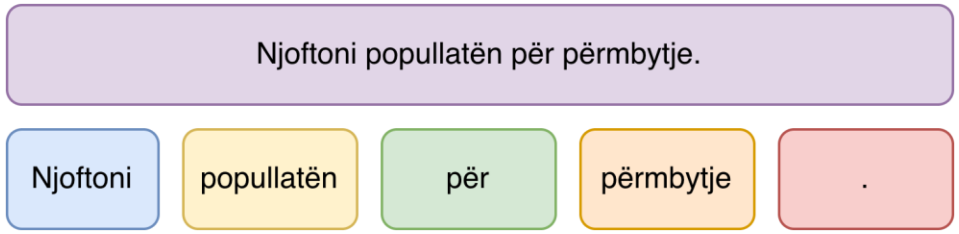
Ndan tekstin në token duke përdorur rregulla të caktuara në mënyrë manuale, zakonisht të bazuara në hapësira, shenja pikësimi, dhe patterns të rregullt.

Karakteristika:

- E fokusuar në linguistikë dhe jo në statistikë.
- E lehtë për tu implementuar dhe personalizuar.

Kufizime:

- Shume specifik për gjuhë të caktuara.
- Nuk mund të përballojë paqartësitë apo variacionet.



Si e shohin MMGj-të tekstin?

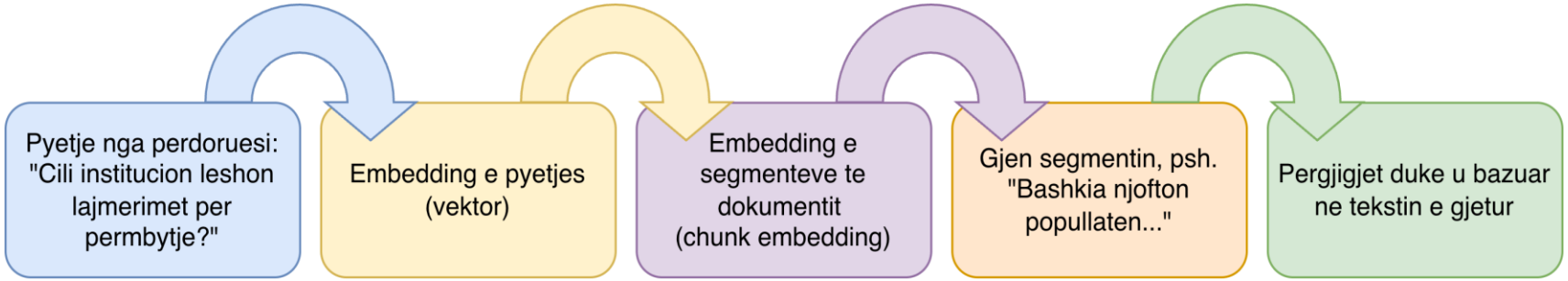
- MMGj-të shohin sekuenca ID-sh të tokeneve, ku çdo ID tokeni lidhet me një vektor brenda një matrice shumë të madhe vektorësh.
- Embeddings janë vektorë shumë-dimensional (psh. 384 dimensione) të cilët përfaqësojnë kuptimin e tokens ose fjalive.
- Dy embeddings janë të ngjashëm nëse paraqiten shpesh në kontekte të ngjashme.

Termi 1	Termi 2	Ngjashmëria
“evakuim”	“strehim”	0.79
“tërmet”	“lëkundje”	0.84
“përmbytje”	“vërshim”	0.81
“njoftim emergjent”	“alarm publik”	0.76

Rëndesia e embeddings në RAG

Embeddings mundësojnë:

- Përputhjen e pyetjeve me segmentet me rëndesi të dokumentit.
- Kërkimin semantik në vend të kërkimit të fjalëve kyçe
- Reduktimin e halucinacioneve
- Garantimin e saktësisë



Segmentimi i dokumentit

Pse segmentim dhe jo gjithë dokumenti?

- Embedding për gjithë dokumentin është tepër i madh.
- Embedding për një fjali të vetme është tepër I vogël.

- Praktika të mira për segmentim: 200–500 tokena
- Duhet respektuar ndarja mes paragrafëve dhe koherenca e informacionit.
- Mund të bëhet mbivendosje 10–15% mes segmenteve.
- Segmentimi I keq → tërheqje e keqe e informacionit

AI Mbrojtja Civile

Embedding ndihmon ne gjetjen e :

- Paragrafit qe diskuton evakuimet
- Zones ku qartesohen pergjegjesite
- Zones ku shpjegohen parametrat e emergjences

AI Mbrojtja Civile

2.0 KUADRI TEKNIK PËR MENAXHIMIN E RREZIKUT

Menaxhimi i emergjencave kërkon një qasje të integruar ndërinstitucionale, ku planifikimi paraprak, reagimi i koordinuar dhe rikuperimi i strukturuar luajnë rol kyç në mbrojtjen e jetëve, pronës dhe mjedisit. Sipas Protokollit Kombëtar të Emergjencave Civile, njësitë vendore dhe rajonale janë përgjegjëse për identifikimin e rrezikut, zbatimin e masave parandaluese dhe reagimin e menjëhershëm në rast të shpalljes së gjendjes së jashtëzakonshme.

2.1 Vlerësimi i rrezikut hidrometeorologjik

Vlerësimi teknik fillon me analizën e të dhënave hidrologjike dhe meteorologjike të mbledhura nga stacionet sinoptike. Parametrat kryesorë përfshijnë:

- nivelin e lumit dhe kapacitetin e shtratit,
- shpejtësinë e rrymës,
- reshjet kumulative në 24, 48 dhe 72 orët e fundit,
- koeficientin e ngopjes së tokës me ujë,
- indeksin e erozionit.

Në rastet kur reshjet tejkalojnë pragun kritik, llogaritja e rrjedhjes së sipërfaqes dhe modelimi hidrodinamik bëhen të domosdoshme për të parashikuar zonat e mundshme të përmbytjes. Njësia e Monitorimit përdor harta dixhitale GIS për të identifikuar habitatet e rrezikuara, infrastrukturën kritike dhe popullsinë vulnerable.

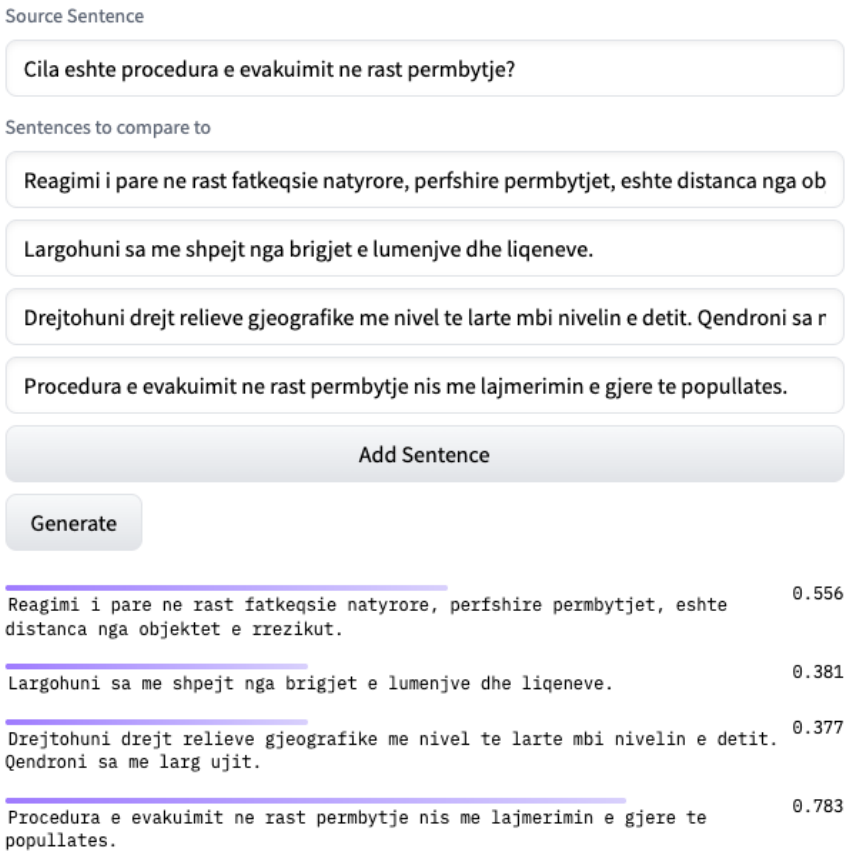
Si embeddings mundësojnë kërkimin semantik?

Kërkim semantik = kërkim i bazuar në kuptim

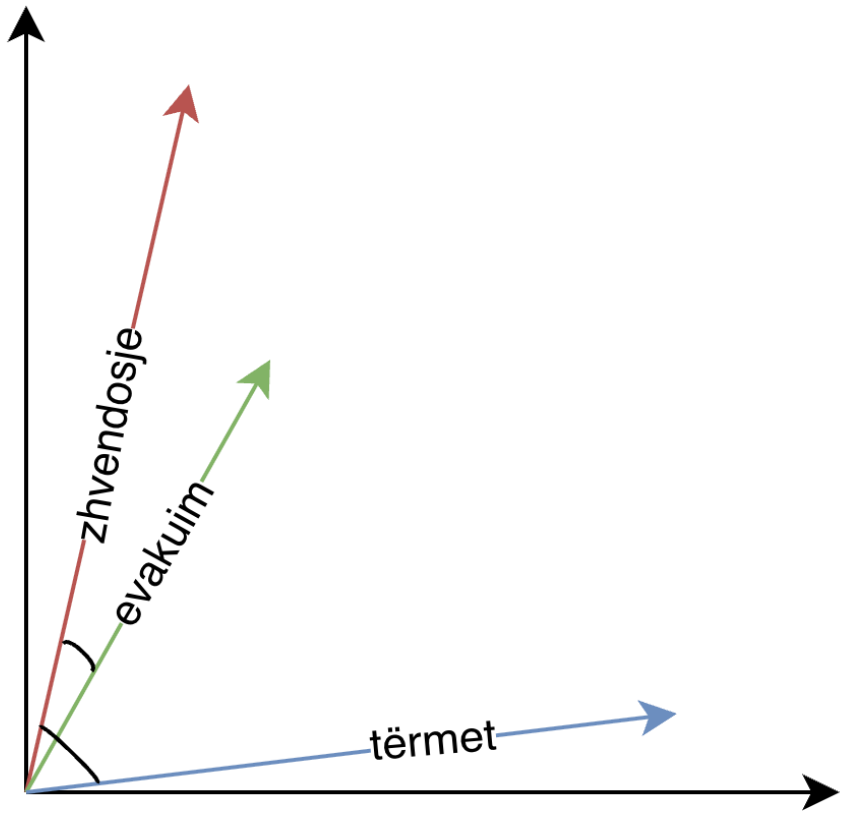
Kërkimi me embedding arrin të gjejë:

- Informacione me rëndesi edhe nëse nuk ka fjalët kyçe.
- Trajton mirë sinonimet.
- Trajton mirë parafrazimet.
- Suporton shumë gjuhë.

Ngjashmëria e embeddings



Matja e ngjashmërisë: Ngjashmëria e kosinusit



Ngjashmëria e kosinusit:

- Mat këndin mes dy vektorëve.
- Ka vlera nga -1 ne 1.
- Sa më e lartë vlera, aq më të ngjashëm vektorët.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Debugging Embeddings: Diagnostikimi i informacioneve jo të sakta

Probleme të zakonshme:

- Segmentim jo i mirë
  - Segment i madh -> zbehet kuptimi
  - Segment i vogël -> humbet konteksti
- Informacione të ndryshme të mbledhura në një paragraf
- Trajtimi jo i mirë i tabelave teknike
- Gabime tokenizimi për gjuhën shqipe

Si të zgjedhim modele për embedding

Modele për të konsideruar:

- all-MiniLM-L6-v2 (I vogël, i saktë, i shpejtë, shumëgjuhësh)
- BGE Base (i saktë por më i avashtë)
- Instructor-Large (kërkon specifikim të detajuar tëdetyrave, punon mirë me taske të veçanta)
- GTE-small (efiçent)

Kritere për zgjedhjen:

- Shpejtësia
- Qëndrueshmëria në domen
- Kërkesat shumëgjuhëshe
- Kufizimet e burimeve

Demo – AI Mbrojtja Civile

- Tokenizojmë me metoden SentencePiece 10 terma emergjence (https://huggingface.co/google/mt5-small).
- Gjenerojmë embeddings për fjalët (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2).
- Zgjedhim një paragraf me 5000 fjalë dhe e ndajmë në segmente (në nivel fjale).
- Krijojmë embeddings duke përdorur modelin MiniLM (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2) (në nivel nënfjale).
- Llogarisim ngjashmërinë.
- Identifikojmë një segmentim të gabuar dhe e rregullojmë.

Çfarë do të trajtojmë në vazhdimësi?

1. ✅ Hyrje
2. ✅ Tokenizimi & Embeddings
3. Arkitektura Transformer
4. Bazat e të dhënave vektoriale (Vector Databases)
5. RAG (Gjenerim i përforcuar nga kërkimi)
6. Promptimi
7. Agjentët (Agents)
8. Përshtatja e modelit (Fine-Tuning)
9. Vlerësimi (Evaluation)
10. Siguria & Përafrimi (Safety & Alignment)
11. Vendosja e modelit (Deployment)
12. Integrimi i Projektit Final

Bibliografia

1. **Sennrich, Haddow & Birch, 2016 — "Neural Machine Translation of Rare Words with Subword Units."**https://arxiv.org/abs/1508.07909
2. **Kudo & Richardson, 2018 — "SentencePiece: A Simple and Language Independent Subword Tokenizer."**https://arxiv.org/abs/1808.06226
3. **Mikolov et al., 2013 — "Distributed Representations of Words and Phrases."**https://arxiv.org/abs/1310.4546
4. **Reimers & Gurevych, 2019 — "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks."**https://arxiv.org/abs/1908.10084
5. **Johnson et al., 2019 — "Billion-scale similarity search with FAISS."**https://arxiv.org/abs/1702.08734
6. **Lewis et al., 2020 — "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks."** https://arxiv.org/abs/2005.11401
7. **Jurafsky & Martin — *Speech and Language Processing* (3rd ed., draft).** https://web.stanford.edu/~jurafsky/slp3/