Gazteak eta euskara sare sozialetan, zer, nori, nork: euskarazko txio formal eta informalak sailkatuz eta konparatuz

Proposer(s) / Proposatzailea(k)

Rodrigo Agerri, Iñaki Alegria

Contact / Kontaktua: email

rodrigo.agerri@ehu.eus i.alegria@ehu.eus

Description / Deskribapena

SARRERA

Gure erlazionatzeko moduak etengabe aldatzen ari dira eta horren erakusle da Informazioaren eta Komunikazioaren Teknologiek eduki duten eragina. Egun, gizarte mendebaldarrean batez ere, pertsona gutxi dira beren eguneroko erlazionatzeko moduan teknologia berriak erabiltzen ez dituztenak. Gazteek bereziki, beren sozializaziorako sare sozialak erabiltzen dituzte, Whatsapp, Youtube, Instagram, Twitter, Facebook eta abar luze bat. Gazteak izango dira ikertuko den unibertsoa, sare sozialen erabiltzaile ohikoenak izateaz gain, etorkizuneko gizartearen isla ere direlako. Bestalde, euskarazko komunitatean zentratzea erabaki da, hizkuntza gutxitu honek XXI. mendeko erronkei nola erantzuten dion aztertzeko asmoarekin.

Twitter izango da lan honetan aukeratuko den sare soziala, beren ezaugarriengatik informazioa ia kasu gehienetan publikoa baita. Informazioa lortzeko erraztasun horrek aukera ematen digu publikatutako testu zati txikien bolumen haundi bat eskuratu eta aztertzeko. Euskara eta gazteak ikertzeko asmoarekin, publikatzen diren txio guzti horien artean euskarazkoak aukeratu beharko ditugu eta euskarazko txio horiek gazte eta heldu artean sailkatu. Ataza horretarako txioen sailkatzaile ez-gainbegiratu bat sortzeari ekingo zaio, euskarazko txio formal eta informalak desberdintzen dituena. Honekin batera, txio informalen corpus bat sortzea izango da beste helburuetako bat, euskarazko idazkera kaletar edo herrikoiagoa nola egiten den erakutsiz. Horrez gain gazteen zein helduen harremantzeko moduak eta gaiak alderatzeari ekingo zaio.

Goals / Helburuak eta Hipotesiak

HELBURUA

1- Sailkatzaile ez-gainbegiratu bat sortzea, euskarazko txio formal eta informalak desberdintzen dituena.

AZPI-HELBURUAK

a) Erabiltzaile gazteak (eta helduak) identifikatu, txio informalen kontzentrazioan oinarrituta.
a1) Gazteen eta helduen harremanak zein gaiak konparatu.
a2) Gazteen eta helduen idazkerak alderatu.
b) Corpus bat osatzea euskarazko txio informalekin.

HIPOTESIAK

1- Gazteek zein helduek hainbat ezberdintasun dituzte, haien artean txioak idazteko modua.
2- Gazteek identifikatuko ditugu txio informalen kontzentrazioan oinarrituz.
3- Gazteen eta helduen harremanak eta gaiak ezberdinak dira.

Requirements / Betebeharrak

Framework / Esparrua

Tasks and plan / Atazak eta plana

DATUEN USTIAKETA

Behin Twitter sare soziala aukeratuta, bertatik datuak ustiatzeko hurrengo metodoak erabili ditugu. Datuen erauzketa burutu ahal izateko Twitterreko APIa erabiliko da, honetarako Pythoneko *tweepy* paketea erabiliz. Datuen bilketa hau burutu ahal izateko hainbat modu ezberdin daude, hiru nagusi ezberdinduko dira hurrengo lerroetan eta bakoitzaren ezaugarrien arabera bat edo beste erabiliko da ustiatuko ditugun datuen arabera.

- *Streaming bidezko erauzketa* : funtzioari termino zerrenda bat pasatzeko aukera dago eta funtzio honek momentu horretatik aurrera txiokatutako elementuak pasatuko dizkigu, termino zerrenda horietako bat topatzen baldin badu.
- *Termino bidezko erauzketa* : termino konkretu bat pasatzen diogu funtzioari eta 100.000 txio inguru lortzeko aukera dago, funtzioa deitzen den momentutik 7 egun atzera eginez maximo.
- *Erabiltzaileen erauzketa* : aurreko bi puntuetan bilaketa termino konkretuen arabera burutzen da, honetan erabiltzaile zerrenda batean oinarrituta, erabiltzaile bakoitzaren “timeline”-a erauzten da. Erabiltzaile bakoitzeko Twitterreko APIak 3200 txioko limitea dauka, beraz erabiltzaile bakoitzeko gehienez kopuru hori lortu ahal izango da.

Datuak erauzteko ikusi ditugun metodo ezberdinetarako, Twitterreko APIak aukera ematen digu hainbat filtro aplikatzeko, guk iragazki zehatz batean jarriko dugu enfasia, hizkuntzan. Modu honetan APIak aukera ematen digu txioak hizkuntzaren arabera aukeratzeko, euskarazko txioen lorpenerako aukera paregabea izanik, asmatze tasa oso ona daukalarik.

Behin euskarazko txioak identifikatzeko eta lortzeko modua ezagututa, euskal txiolariak identifikatzeari ekingo diogu, horretarako bi pausu nagusi emango ditugularik:

1. Lehenengo pausua euskaraz argitaratzen diren txioak jasotzea izango da, horretarako termino bidezko erauzketa erabiliko dugularik. Ataza honetarako, euskaraz gehien aipatzen diren hitzak erabiliko ditugu, “stopword” delakoak. Hitz hauen artean *dira, eta, egin* etab. erabiliko dira besteak beste. Gero, asmatze tasa hobetzeko asmoarekin, lortutako txio hauek hizkuntzaren arabera filtratuko dira, euskarazko txioak lortuz. Era honetan, txio hauek bidali dituzten erabiltzaileak zerrenda batean gordeko dira.
2. Bigarren pausu moduan, euskarazko txioak argitaratu dituzten erabiltzaile guztien analisia burutuko dugu. Honela, erabiltzaile bakoitzaren azkeneko 200 txioak jasoko dira eta hauen % 20a euskaraz izatekotan, euskal txiolaria dela onartuko da. Era honetan, euskal txiolarien zerrenda bat lortuko da eta hauen *timeline* a ustiatu erabiltzaileen ustiaketa funtzioa erabiliz. Behin euskal txiolarien *timeline* a ustiatuta, datu basea sortu ahal izango da eta txioen sailkapenarekin hasi ahalko da.

DATUEN ANALISIA

Datuen analisia burutzerako orduan entropia eta perplexitya erabiliko dira, txio informalak detektatzeko asmoarekin. Horretarako, euskara formaleko corpus bat erabiliko da eta neurriek adierazten dutenean determinatu egingo da txio formala edo informala den.

Modu berean, txio bakoitzaren identifikazioan (formal/informal) oinarrituta, erabiltzaileak idatzitako txioen batezbesteko bat egingo da, erabiltzailearen tipologia zehazteko asmoarekin. Honela, erabiltzaileak sailkatu ahalko ditugu, gazte eta heldu ezberdinduz idazteko moduan oinarrituz.