Donostia/San Sebastián 2016 “5.2.8 DSS2016 Behagunea - Hirikia kaia” web aplikazioa

Proiektuaren Memoria
Elhuyar Hizkuntza eta Teknologia

Proiektu arduraduna (Elhuyar): Iñaki San Vicente Roncal	DSS2016ko arduraduna (Hirikia kaia): Aitzol Astigarraga

2016eko maiatzaren 10a

Zelai Haundi, 3. Osinalde Industrialdea. 20170 Usurbil. Gipuzkoa. tel.: 943363040. h.el.: i.sanvicente@elhuyar.com

Aurkibidea

Zelai Haundi, 3. Osinalde Industrialdea. 20170 Usurbil. Gipuzkoa. tel.: 943363040. h.el.: i.sanvicente@elhuyar.com

1 Sarrera

Donostia/San Sebastián 2016 Fundazioak (aurrerantzean, DSS2016EU) DSS2016ko ekitaldien inguruan herritarrek idatziko dituzten iritziak automatikoki aztertzen dituen web aplikazioa garatzeko Elhuyar Fundazioaren zerbitzuak kontratatu ditu.

Txosten honek garatutako plataforma digitalaren ezaugarriak eta egindako lanaren inguruko xehetasunak ematen ditu. Garatutako soluzioa ikusgai dago http://behagunea.dss2016.eu/ helbidean. Orokorrean, "Behagunearen" helburua Donostia 2016 ekimenaren markoan antolaturiko ekitaldiek gizartean duten oihartzuna aztertzea da, ondorengo ezaugarri nagusiak dituelarik:

- Sistemak DSS2016 ekimenaren barruan ematen diren proiektuen inguruko iritzia jasotzen du, denbora errealean. Iritzi bat publikatu eta handik 10 minutura iritzi horiek linguistikoki prozesatuta eta polaritatea aztertuta "Behagunean" ikusgai daude.
- Aipamenak hainbat iturritatik eskuratzen ditu aplikazioak horretarako sortutako berariazko softwareari esker. Egunkariak, sare sozialak, eta blog-ak ustiatzeko gai da.
- Iritzien analisi aurreratua. Sistemak mezuen polaritatea detektatzen du, mezuetan aipatzen diren gaiak eta haien egileak identifikatzen ditu. DSS2016ren domeinura egokitutako iritziak aztertzeko teknologia garatu da, 4 hizkuntzatarako, Euskara barne.
- Datuen bisualizazio aurreratua: Sistemak DSS2016 ekimenaren barruan ematen diren proiektuen inguruko iritzia jasotzen du,

Zelai Haundi, 3. Osinalde Industrialdea. 20170 Usurbil. Gipuzkoa. tel.: 943363040. h.el.: i.sanvicente@elhuyar.com

denboran zeharreko aldaketak erakusten ditu, egile aktiboenak identifikatzen ditu, eta jasotako datuen gaineko estatistikak sortzen ditu, datu guztiak modu erosoan kontsultatzeko bisualizazio interaktibo aurreratua eskaintzen dituelarik.

- Zaintza sistemaren kudeaketa aurreratua: Kudeatzaileak aukera du monitorizatu nahi diren gaiak zein iturriak konfiguratzeko, eta estatistika atal berezitua ere badu. Horretaz gain, jasotako iritzien eskuzko berrikuspena egiteko aukera ematen zaio.

Sistema irekien filosofiarekin guztiz bat datorren soluzio teknologikoa garatu da. Proiektuan zehar garatu diren teknologia guztiak software librea erabiliz egin dira, eta lizentzia libreekin eskuragarri daude sarean.

Txosten hau honela antolatzen da hemendik aurrera; hurrengo atalean Elhuyar Fundazioak garatutako soluzio teknologikoa deskribatzen da. 3. atalak 2016 ondorenerako proiektuan dituen jarraipen asmoak aurkezten dira. 4. atalak proiektuaren eta garatutako teknologiaren ebaluazioa egiten da.

2 Web aplikazioa – Soluzio teknologikoa

2.1 Sarrera

Aurreko atalean azaldu dugun bezala, proiektu honen helburua aplikazio bat garatzean datza, zeinaren bitartez automatikoki analizatuko diren Donostia 2016 ekimeneko ekitaldien inguruan herritarrek idazten dituzten mezuak, denbora errealean. Testu horiek linguistikoki analizatzen ditu, eta bertan azaltzen diren iritzien sailkapena burutzen du, zein gairi dagozkion identifikatuz eta iritzi horiek positibo, negatibo edo neutralak diren esanez.

Datu horiek webgunean publikoki ikusgai daude, modu erakargarrian, erabiltzaileei nabigazio eroso eta intuitiboa eskainiz eta datuen kontsulta aurreratua ahalbidetuz.

Teknologiari dagokionez, iritzi bat adierazten duen testua aztertzeko orduan, gai izan behar dugu, testuan adierazi diren elementu positiboak edo negatiboak identifikatzeko, eta iritzia zeri buruzkoa den detektatzeko, edo iritzia nork adierazi duen esateko. Horretarako, hizkuntzaren prozesamendurako teknologiak erabiltzea ezinbestekoa da, testuaren analisi egoki bat. Gure proposamenak Ixa-Pipes tresna linguistikoak erabiltzen ditu oinarrizko analisia burutzeko, EHUko IXA taldeak garatuak. Baina ez hori bakarrik, informazioaren erauzketa ere egin behar da, iritzien erauzketa. Horretarako erregela eta ikasketa automatikoan oinarritutako algoritmoak, artearen egoerakoen parekoak, erabiltzen ditu gure proposamenak.

Garatutako soluzioa, domeinura egokitutakoa da,, izan ere, emaitza domeinu kulturalera eta euskal herriko egoera soziolinguistikora

moldatutako sistema bat da, hots DSS2016ren beharretara egokitutako sistema.

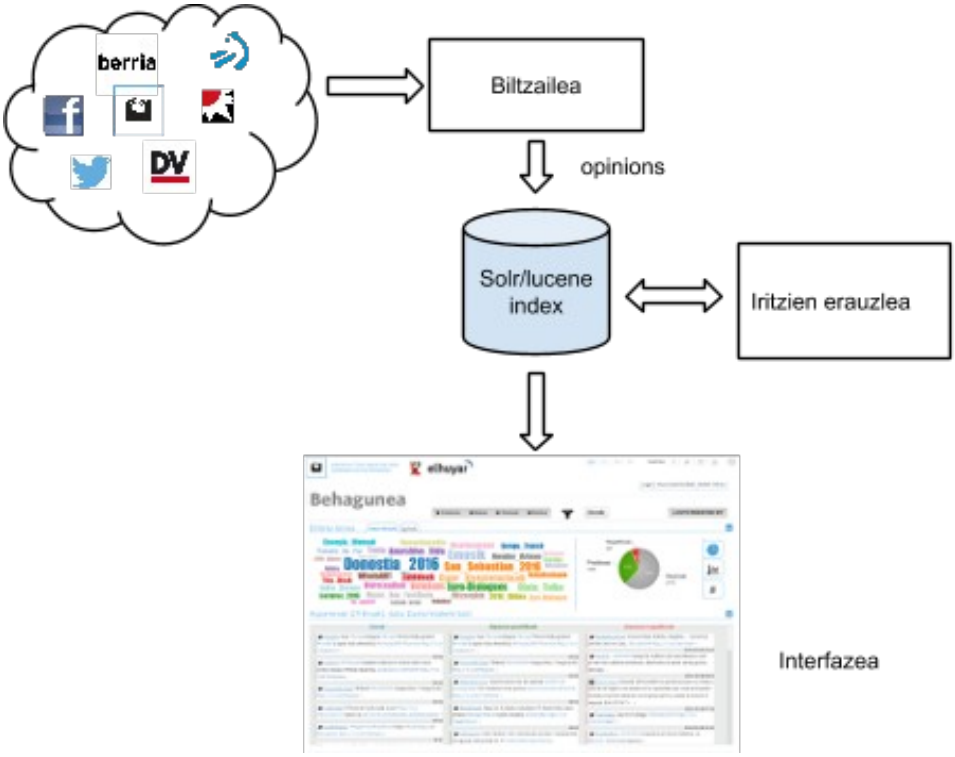
Ondorengo ataletan garatutako aplikazioaren ezaugarri zehatzak aurkezten dira, eta erabiliko den teknologiaren inguruko xehetasunak ematen dira.

2.2 Sistemaren arkitektura

Web aplikazioa MVC (Model–view–controller) arkitektura batean oinarrituta dago. Arkitektura honek datuak eta negozioaren logika erabiltzailearen interfaze eta gertaerak kudeatzeko moduluetatik bereizten ditu, etorkizuneko garapenak eta mantentze lanak erraztuz. Zehazki, Django WAFa (Web Application Framework) erabili da interfaze nagusia garatzeko eta datu-biltegiarekin komunikazioa inplementatzeko..

Sistema hiru modulu nagusiz osatua dago:

- Biltzailea (crawler-a): Iritzien bilketarako moduluak interneteko iturri ezberdinetatik jasoko ditu edukiak biltegi batean gordetzeko.
- Iritzien erauzlea: Prozesamendu kateak jasotako edukiak prozesatu eta bertako iritzien analisia egiten du. Erauzle honek prozesamendu linguistiko osoa burutzeko katea integratzen du, lematizazio, etiketatze linguistiko eta polaritate sailkatzailea barne hartzen dituelarik.
- Iritziak aztertzeko interfazea: Bildutako testuetan aurkitutako iritziak modu grafikoan kontsultatzeko funtzionalitateak eskaintzen ditu, bilketa moduluaren kudeaketa egitea ahalbidetzen du, eta sailkapen automatikoaren emaitzak berrikusteko aukera ematen du.



Irudia 2: Proposatutako sistemaren arkitektura, MVC patroian oinarritutakoa.

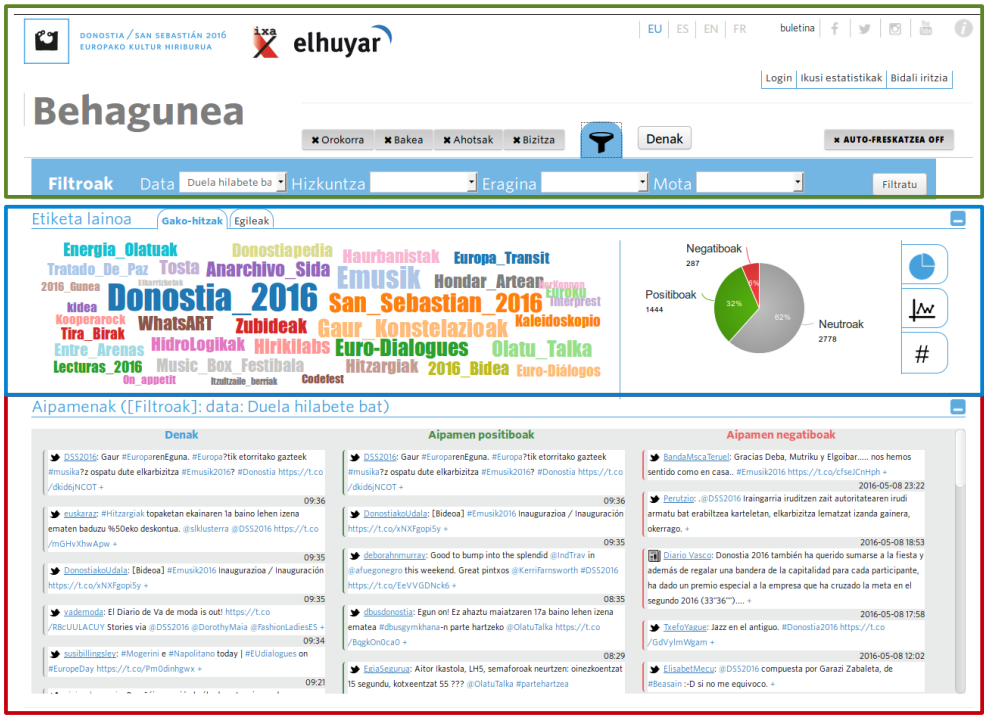
2.3 Erabiltzailearen interfazea

RWD (Responsive Web Design) filosofian oinarritutako web-interfazea garatu da, tresnaren funtzionalitatea modu intuitiboan erabiltzea ahalbidetzen duena, eta interfazea bera zenbait gailutara (tabletak, smartphone-ak, PCak...) egokitzen dela bermatzen duena. Interfazea eleaniztuna da 4 hizkuntzetan (Euskara, gaztelania, ingelesa eta frantsesa).

Ondoren erabiltzailearen interfazearen pantaila nagusiak eta eskaintzen dituen nabigazio aukerak azaltzen dira. Hurrengo ataletan agertuko diren bisualizazio taula eta horien gainean inplementatutako elkarrekintza oro d3.js javascript liburutegiarekiko

2.3.1 Pantaila nagusia, nabigazio orokorra

Pantaila nagusiak DSS2016ko proiektuen uneko argazkia erakusten du, denbora errealean. 3 elementu nagusi daude pantaila honetan: goiburua filtro eta elkarrekintza aukerekin, proiektu zein egileen bisualizazioak, eta txioen erakusleihoa.



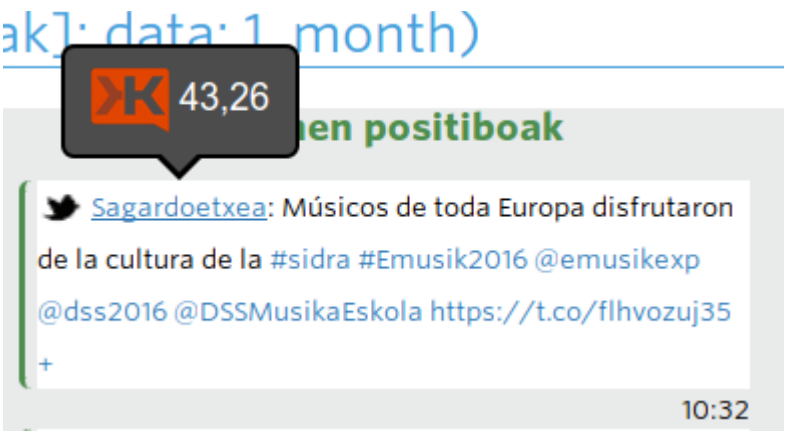
Irudia 3: Interfazearen pantaila nagusia

Elementu nagusia bisualizazioen atalak osatzen du. Bertan, gaien eta iritzi-emaileen lainoak dira elementu garrantzitsuena. Lainoak gai esanguratsuenak eta iritzi-emaile aktiboenak begiratu bakarrean ikusteko aukera ematen du. Modu lehenetsiak gai edo proiektuen lainoa ikusten da, egileen lainoa ikusteko dagokion fitxa aukeratu behar dugularik. Lainoen eskuinaldean, iritzien laburpen atala dugu ikusgai. Bertan, iritzi 3 informazio ditugu eskuragarri: (i) iritzien positibo eta negatiboen oreka

sektore-diagrama baten bitartez, (ii) iritzien eboluzioa denboran zehar erakusten duen grafikoa eta (iii) jasotako iritzi kopuruen inguruko estatistikak.

Irudia 4: Bisualizazioak

Aipamenen erakusleihoan txioak 3 zutabetan agertzen dira. Lehen zutabean azken aipamenak erakusten dira, eta hurrengoetan positibo ala negatibo sailkatutako aipamenak. Iritzi bakoitzaren iturria, iturri horren "ospea", data eta egilea ere erakutsiko dira (ikus irudia 5).



Irudia 5: Aipamen baten ezaugarriak

Elkarrekintza bideratzeko, hainbat filtro eskaintzen dira, goiburuko menuan eskuragarri. Filtroek ondorengo aukerak eskaintzen dituzte:

- Epe ezberdinen arteko iritziak bakarrik ikusi.
- Hizkuntza jakin bateko iritziak ikusi.

- "Ospe" maila batetik gorako egileen iritziak bakarrik ikusi.
- Iturri mota jakin bateko iritziak bakarrik ikusi (prentsa, sare-sozialak).
- Itsasargi zehatz bateko proiektuen inguruko aipamenak ikusi.

Filtro hauetako bat edo gehiago aukeratutakoan pantailako elementu guztien informazioa filtroen aukeraketarekin arabera eguneratzen da, koherentzia mantenduz. Filtro horietaz gain, bai bisualizazioek eta baita aipamenek ere elkarrekintza eskaintzen dute:

- Zerrendako iritzi bakoitza interaktiboa izango da, iritziaren egilea klikagarria izanik. Egilearen gainean klikatuz gero pertsona horrek emandako iritzi guztiak bistaratuko dira. Aipamen bat bere jatorrizko helbidean ikusteko aukera ere badugu "+" ikurra sakatuz.
- Egile eta gaien lainoetako elementuak ere interaktiboak dira, eta filtro modura funtzionatzen dute, hots, elementu bat klikatzerakoan, interfazea freskatu egiten da, elementu horrekin lotutako informazioa erakutsiz.

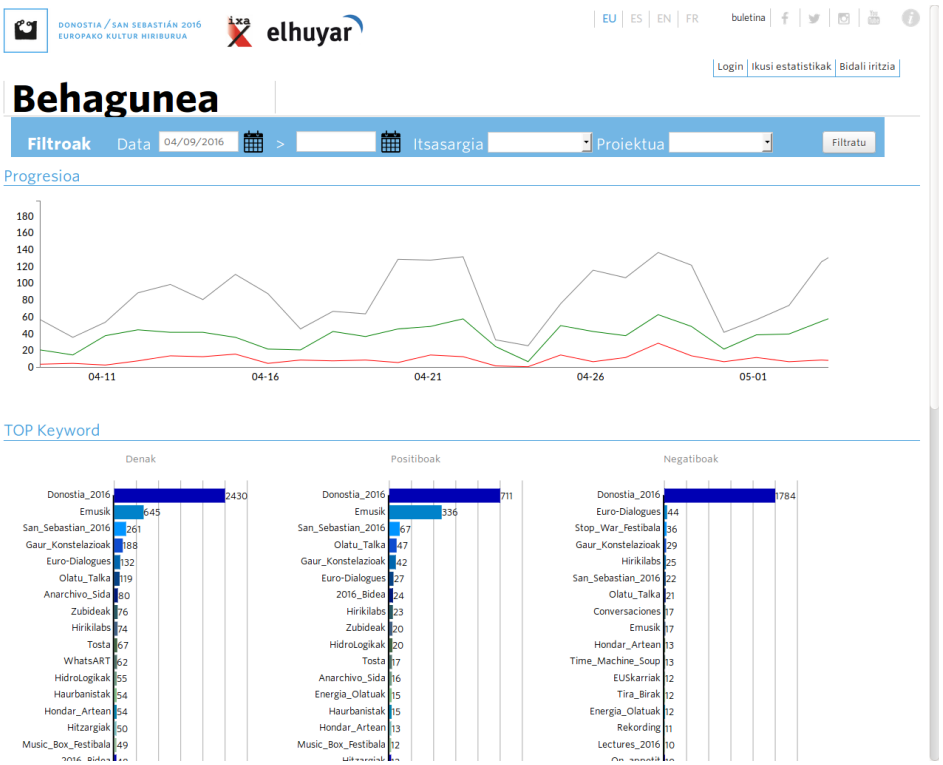
Azkenik, beste bi elementu aipatuko ditugu. Batetik freskatze automatikoa ahalbidetzeko botoia. Honekin pantailako informazioa 15 minutuero automatikoki eguneratzen da. Horrela "Behagunea" pantaila nagusi batean eduki liteke eta informazio eguneratua ikusi uneoro. Bestetik "Iritzia bidali" aukeraren bidez edozein erabiltzailek bere iritzia emateko aukera du, zuzenean plataformaren bitartez.

2.3.2 Estatistikak

Goiburuko menuan "Ikusi estatistikak" klikatuta, estatistika orokorren atala zabalduko dugu. Bertan iritzien eboluzioaren grafikoa aurkituko

dugu lehenengo gauza, eta ondoren, gai aipatuenen, eta, prentsako iturrien eta Twitterreko egileen rankingak ikus daitezke. Bloke bakoitzean 3 ranking erakusten dira. Lehenengoan elementu aipatuenak eta bakoitzak jasotako aipamen kopurua agertzen da. Erdiko rankingak aipamen positiboei dagokien informazioa ematen du, berriz ere elementu aipatuenak eta jaso dituzten aipamen positibo kopurua, eta azken zutabean aipamen negatiboei dagozkienak. Egileei dagozkien rankingen kasuan, egile bakoitzak botatako aipamen kopuruak erakusten dituzte rankingek.

Pantaila nagusiaren antzera, informazioa filtratzeko aukerak ere eskaintzen dira. Kasu honetan, data zehatzen arteko bilaketak egin ahal izango ditugu, proiektuaren edo itsasargien araberakoak.

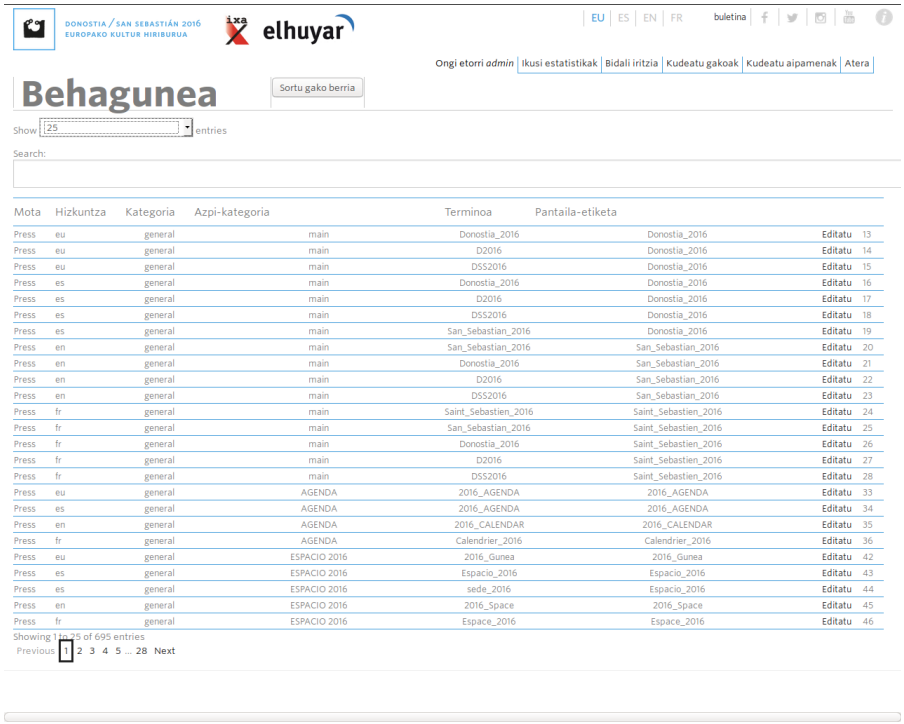


Irudia 6: Estatistika orokorrak, iritzien denboran zeharreko eboluzioaren grafikoa eta gai aipatuenen rankingak ikusten dira.

2.3.3 Kudeatzailearen aukerak

Behaguneak kudeatzaile rol bat ere inplementatzen du. Kudeatzaile horrek funtzionalitate aurreratuak ditu eskura, batetik, datu-bilketaren ezarpenak aldatu ahal izateko, eta, bestetik, jasotako datuen gaineko kontrola eramateko. Ondoren kasu bakoitzean ematen diren aukerak azaltzen ditugu.

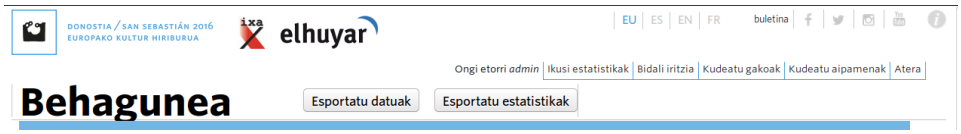
- Gakoen kudeaketa: aipamenak aurretik definitutako gako-hitz zerrenda baten arabera biltzen dira. Kudeatzaileak zerrenda hori ikusi eta editatzeko aukera du.



Irudia 7: Gakoak kudeatzeko intefazea

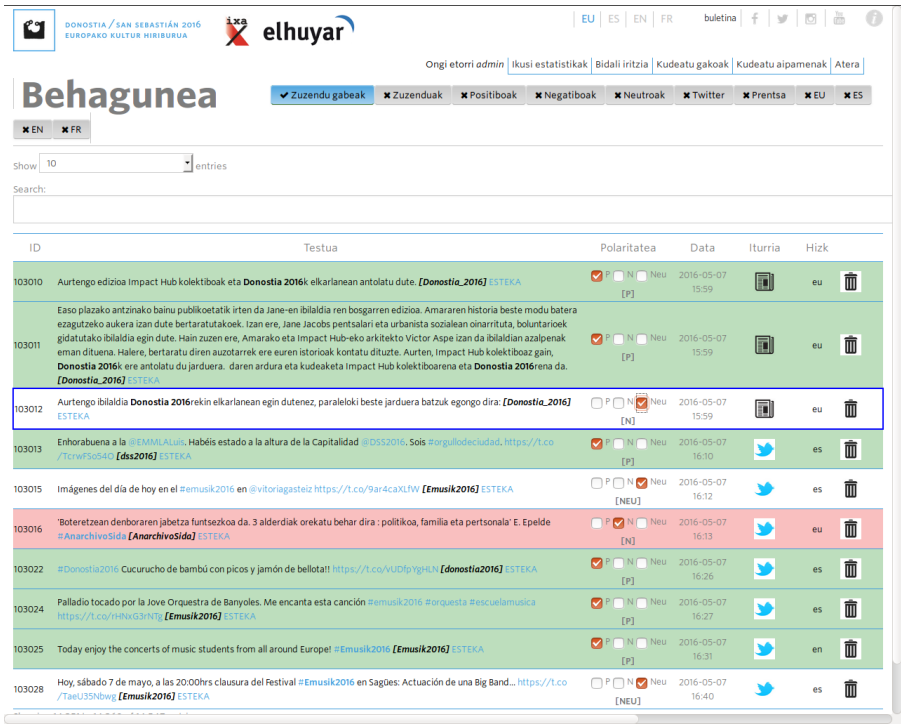
Gako bakoitza kategoria eta proiektu batekin lotuta dago. Horrela, posible da proiektu edo ekimen berriak jarraitzen hastea, dagoeneko jarraitzen ari garen ekimenei gako-hitz berriak gehitzea edo daudenak kentzea.

- **Estatistika aurreratuak:** estatistiken atalean kudeatzaileak bi aukera gehigarri ditu eskuragarri, erabiltzaile arruntekin alderatuta: “Esportatu datuak” eta “Esportatu estatistikak”. “Esportatu datuak” aukerak



Irudia 9: Kudeatzailearen aukera osagarriak estatistiken atalean

- **Zuzenketa:** Iritziaren azterketa automatikoa izanik, demagun erabiltzaile gisa oker sailkatutako iritzi bat topatzen dugula. Kudeatzaileak aukera du jasotako iritziak edozein unetan errepasatu, eta hala beharrezko ikusten badu horiek polaritatea zuzentzeko klik soil bat eginez. Aipamenak ezabatzeko aukera ere badu kudeatzaileak, adibidez, gaikoa ez den aipamen bat antzematen bada, edo inolako iritzirik adierazten ez duten spam mezuak detektatzen dituenean. Honek iritzi erauzketa hobetzen lagunduko digu epe ertainera, sistemak bere erroreetatik ikasi ahal izango baitu.



Irudia 9: aipamenak kudeatzeko atala. Aipamenak bere polaritatearen araberako koloreaz erakusten dira. Aipamen baten polaritatea zuzentzeko polaritate zuzena markatu besterik ez dago. Zuzendutakoak urdinez nabarmentzen dira.

2.4 Iritzi bilketa: Iturriak

Behatokian iritziak biltzeko hautatutako edukiak DSS2016EUrekin batera landu dira. Ondoren jarraitzen diren iturrien zerren da ematen da:

- Sare sozialak:
    - Twitter
    - Instagram eta Facebook biltzea ere aztertu da, baina lehena baztertu egin da testuzko eduki minimoa izanik, iritziak jasotzeko zailtasunak planteatzen dituelako. Facebook-i dagokionez, erabiltzaileek mezu positiboak botatzeko joera nabarmena dute, zenbait ikerketek erakutsi duten bezala, mezu positiboek negatiboek baino

eragin handiagoa dutelako, jarraitzaile gehiago lortu araziz. Ondorioz Facebook ez erabiltzea erabaki da.

- **Prentsa digitala:** Ondorengo taulan zehazten dira biltzen diren hedabide digitalak:

BERTAKOAK	
Hedabidea	**Hizkuntza**
DEIA	es,eu
Gara	eu,es,fr
Hitza	
Busturialdeko Hitza	eu
Goierriko Hitza	eu
Irutxuloko Hitza	eu
Lea-Artibai eta Mutrikuko Hitza	eu
Oarso Bidasoko Hitza	eu
Bidasoko Hitza	eu
Oarsoaldeko Hitza	eu
Tolosaldeko Ataria	eu
Urola Kostako Hitza	eu
Argia	eu
EL CORREO	es,eu
Info 7	es,eu,fr
Diario Vasco	es,eu
Radio Vitoria	es
Artezblai	es
Cuatro	es
Onda Vasca	es
Berria	eu
Goiena	eu
Noticias de Gipuzkoa	es,eu

AGENCIA EFE	es
EL PAIS (CULTURA)	es
EL PAIS SEMANAL	es
TV3	es,eu
Radio Donosti	es
Europapress	es
Eitb	es,eu
EUSKADI IRRATIA	eu
EL MUNDO	es
Teledonosti	es,eu
Hamaika tv	es,eu
Gaztezulo	eu
Naiz	es,eu,fr
RTVE	es
ABC	es
DIARIO DE NAVARRA	es
Radio France	fr
France Bleu	fr
El Periódico de Catalunya	es
La Vanguardia	es
NAZIOARTEKOAK	
Ouest-France	fr
Reuters Breaking Views	fr
Le Monde	fr
Liberation	fr
France 24	fr
Le soir	fr
Financial Times	en
Le Figaro	eu
The Times	en
L'Express	fr
The Economist	en
CNN	en
Arte tv	fr

Zelai Haundi, 3. Osinalde Industrialdea. 20170 Usurbil. Gipuzkoa. tel.: 943363040. h.el.: i.sanvicente@elhuyar.com

2.5 Aipamen biltzailea: MSM

Iritzien behatoki bat martxan jartzeko lehen pausua sarean sakabanatuta daude iritziak atzeman eta horiek biltzea da. Baina ez hori bakarrik, gure kasuan prozesu hori iritziak argitaratu eta ahalik epe laburrenean burutzea da helburua, segundo edo minututako atzerapenaz.

Azkartasun hori ez da iturri guztien kasuan berdina. Sare sozialen kasuan edozein unetan sor daitezke mezu berriak, eta beraz etengabe “entzuten” egotea eskatzen dute, eduki berriak noiz agertuko adi. Aldiz prentsa digitalean iritzi artikuluak sortzeko maiztasuna askoz ere baxuagoa da, eta ondorioz nahikoa izan daiteke egunean behin eduki berririk dagoen aztertzearekin. Gauza bera gertatuko litzateke blogekin, izan ere ezohikoa da astean 2-3 artikulu baino gehiago idaztea.

Aurre egin beharreko beste arazo bat edukia atzitzeko modua da. Sare sozialak, Twitter edo Facebook kasu, enpresa pribatuak dira, bere edukia API bidez jartzen dute eskuragarri, eta beraz, sare bakoitzarentzat ad-hoc prestatutako bilketa tresnak garatu behar dira, haiek jarritako baldintzak betez. Blog eta prentsaren kasuan, edukiak RSS bidez banatu izan dira. Azken aldian, zenbait komunikabide, Twitter bidez zabaltzen hasi dira haien eduki guztiak.

Aurrekoak kontutan izanik, “DSS2016 Behatokian” bi estrategia baliatzen dituen sistema garatu da: Media Source Monitor (MSM). MSM Javaz inplementatu da, eta hainbat funtzionalitate eskaintzen ditu, aipamenen bilketa, egileen eragin mailaren kalkulua, eta datu-biltegiarekin komunikazio barne:

- Twitter bezeroa: Public Streaming API delakoarekiko lotura sortzen du, APIra konektatzerakoan informazio geografikoaren, gako-hitzen, erabiltzaileen eta hizkuntzaren araberako filtroak ezartzeko aukera ematen duelarik. Konexioa irekita mantentzen du etengabe, edukia sortu ahala jasotzen delarik. Konexioa eteten bada konexioa berrabiarazten du automatikoki.
- RSS biltzailea: Prentsa digitaleko eta blogetako edukien bilketa burutzeko RSS jarioetan oinarritutako biltzailea inplementatzen du MSM-k. RSSak aldizka irakurri behar dira eduki berririk iritsi ote den aztertzeko. Erabiltzailearen esku uzten da maiztasuna erabakitzea.
- Egileen ospea: Egile zerrenda bat emanda, MSM-k egile bakoitzaren "eragin maila" edo "ospea" kalkulatzen du. Egile motaren arabera indize ezberdinak erabiltzen dira:
    - Klout: Twitter erabiltzaileen ospea neurtzeko Klout[1] API-a erabiltzen da.
    - Ahrefs: Web domeinuen ospea neurtzeko Ahrefs[2] indizea erabiltzen da, DomainStats.io API[3] aren bitartez.

MSM-k konfigurazio fitxategi baten bitartez funtzionatu dezake, eta emaitzak irteera estandarrera bideratu, baina Behaguneako datu-biltegiarekin komunikazio zuzena ere badu. Horrela, parametro guztiak datu-biltegitik irakurtzeko gai da, eta jasotako edukiak datu-biltegira bideratzeko ahalmena du.

[1] https://klout.com/s/developers/v2

[2] https://ahrefs.com/

[3] http://api.domainstats.io/

2.6 Iritzien analisia: prozesamendu kate eleaniztuna

Jasotako mezuetan dauden iritzien analisia egiteko, Elhuyar Fundazioak garatutako iritzien analisirako softwarea erabili da, EliXa (San Vicente et al., 2015) izenekoa. Softwarea hau Javaz inplementatuta dago eta egungo bertsioa GPL lizentziapean banatuko da. Software hau OpeNER proiektu Europarrean (FP7) eta Ber2tek proiektuan (Etortek IE12-333) garatutako teknologiaz baliatuko da.

Dataset	Aipamenak	P	N	NEU
eu	2937	931	408	1598
es	4754	1487	1303	1964
en	12,273	4,654	1,837	5,782
fr	11,071	3,459	2,618	4,994

Taula 1: Osatutako bildumen inguruko estatistikak.

EliXak ikasketa automatikoa erabiltzen du polaritatea sailkatzeko, polaritate lexikoen ezagutzarekin konbinatuta. Hurbilpen gainbegiratua dugu hau. Izan ere, sistema aurretik sailkatuta dauden adibideetatik eredu bat ikasten du, eta ondoren testu berria datozenean eredu horren arabera sailkatzen ditu. Hurbilpen hau erabiltzeko, beharrezkoa da ereduak ikastek sailkatuta dauden adibideak edukitzea. EliXaren arkitekturak ereduak ikasteko orduan hainbat ezaugarri gehitu edo kentzeko aukera ematen du, eta baita ikasitako ereduak ebaluatzeko ere. Horrela, uneko domeinura optimizatutako ereduak sortzea erraza eta eraginkorra bihurtzen da.

EliXak Euskarri bektoredun makina (ingelesez Support Vector Machine, edo SVM) motako algoritmoa erabiltzen du sailkapena burutzeko eta polaritate lexikoetatik erauzitako informazioa ezaugarri linguistikoekin konbinatzen du. Erabiltzen dituen ezaugarri nagusiak hauek dira:

- Polaritate lexikoaren informazioa
- Informazio linguistikoa (adibidez, kategoria gramatikalak, puntuazio markak, ...)
- Egitura sintaktikoak (e.g., “egokia da” vs. “ez da egokia”)
- Informazio estatistikoa (adibidez, testu luzera, hitz positibo/negatibo kopurua, ...)
- Sare sozialekin lotutako informazioa (adibidez, emotikonoak, letra larrien kopurua, ...)
- Hitzen lema n-gramak
- PoS ngram-ak

Behagunea proiektuan, EliXaren hainbat egokitzapen burutu dira. Batetik, euskararako eta frantseserako polaritate sailkatzaileak garatu dira, horrek barne hartzen dituen lanekin: adibide bildumak sortu eta polaritateaz anotatu eta polaritate lexikoak sortu hizkuntza horietan. Bestetik, domeinura egokitutako ereduak sortu dira, horiek sortzeko ezaugarri egokienak zein diren ikertu dugularik.

DSS2016ren aplikazio eremura egokitutako adibideak edo corpusak bildu dira (2015eko urritik 2016ko martxoa bitarte), eta horien polaritatea eskuz etiketatu da, ikasketa automatikoko sistemak entrenatu ahal izateko. Taula 1-ek sortutako datu-bildumen inguruko informazioa ematen du. Ingelesezko (en) eta frantsesezko (fr) bildumen kasuan DSS2016ko informazio nahikorik ez zegoela eta domeinu orokorreko bildumak ere erabili dira adibidea osagarriak lortzeko.

Hizkuntza	doitasuna	fpos	fneg	fneu
eu	72.66	0.646	0.636	0.789
es	70.78	0.669	0.725	0.724
en	68.0946	0.685	0.525	0.722
fr	64.33	0.588	0.605	0.698

Taula 2: Entrenatutako ereduen doitasuna. Azken hiru zutabeek kategoria bakoitzeko lortzen den f1-score balioa ematen dute.

Sortutako sailkatzaileek hizkuntzaren araberako doitasun ezberdinak dituzte. Taula 2 ematen dira sailkatzaile ezberdinen dituzten doitasun mailak.

Oinarrizko prozesamendu linguistikoa: Ixa-pipes

EliXak polaritatearen azterketa egiteko testuak aurretik linguistikoki etiketatu behar ditu, hitzen lemak, kategoria gramatikalak, entitateak eta abar identifikatuta eduki. EliXa aurretik etiketatutako testuen iritzia erauzteko gai da, baina oinarrizko prozesamendu linguistikoa ere integratua du bere baitan, landu gabeko testu gordina aztertu ahal izateko.

Horretarako EHUko IXA taldean garatutako IXA-pipes[4] (Agerri et al., 2014) tresna erabiltzen da. Ixa-pipes erabiltzeko prest datorren hizkuntza naturalaren prozesamendurako tresna multzoa da.

Behagunea proiektuaren aurretik, tokenizazioa (ixa-pipe-tok), lematizazioa eta analizatzaile morfologikoa (ixa-pip-pos), analizatzaile sintaktikoa (ixa-pipe-parse) eta entitate ezagutzailea (ixa-pipe-nerc) eskaintzen zituen gaztelania eta ingeleserako, eta partzialki galiziera ere bai. Gainera, Ixa-pipesen garapena OpeNER[5] proiektuari lotuta dago, eta bertan garatutako prozesamendu katearekin bateragarria da. Horri esker,

[4] http://ixa2.si.ehu.es/ixa-pipes/

[5] http://www.opener-project.eu/

frantsesa, nederlandera, italiera eta partzialki alemana ere prozesatzeko aukera ematen du. “DSS2016 Behagunea” proiektuaren baitan, euskararako eta frantseserako tokenizatzaile eta analizatzaile linguistikoak garatu dira.

3 Software librea, lizentziak

Gure proposamenak software librearen filosofiarekin erabat bat egiten du. 2. atalean aurkeztutako software oro lizentzia libreekin banatutako teknologien gainean garatu da. 3 Taulak garapenean erabilitako osagai nagusien lizentziak eta proiektu honetan duten erabilera laburtzen du.

Softwarea	Lizentzia	"D2016 behatokia" Proiektuan duen eginkizuna	Eskuragarri non
Django Python Web framework	BSD	Interfazearen oinarria	https://www.djangoproject.com/
IXA-pipes	Apache 2.0	Testuen prozesamendu linguistikoa egiteko liburutegiak	http://ixa2.si.ehu.es/ixa-pipes/
Weka	GNU GPL	Ikasketa automatikoa burutzeko liburutegia	http://www.cs.waikato.ac.nz/ml/weka/
EliXa	GNU GPL	Polaritatearen azterketarako modulua.	https://github.com/Elhuyar/EliXa
Apache Solr	Apache 2.0	Datuen indexatzea, bilaketa motorra.	http://lucene.apache.org/solr/
D3.js	BSD	Bisualizaziorako erabilitako Javascript liburutegia	http://d3js.org/

Taula 3: Garapenean erabilitako software nagusiak eta beren lizentziak.

Garatutako softwarearen lizentziari dagokionez, 2 atalean, sistemaren arkitekturan deskribatzean aipatutako hiru moduluak modu independentean banatzen dira, bakoitza bere lizentziapean, baina, beti ere, DSS2016EUren asmoekin bat eginez, lizentzia horiek libreak izanik. Horrenbestez ondorengo lizentziak proposatzen ditugu:

- Interfazea, Django-ren izpirituari jarraiki, BSD lizentziapean banatuko da.
- Datuen bilketarako modulua (MSM) GNU GPL (V3) lizentziapean banatzen da.
- Azkenik, EliXa polaritatearen analisirako moduluak GNU GPL (V3) lizentzia du. Izan ere EliXaren egungo bertsioak GPL lizentziadun liburutegiak erabiltzen ditu, eta honek garatutako softwarea ere lizentzia horrekin banatzera behartzen du.

Garatutako software oro biltegi publikoetan eskuragarri dago jada, ondorengo helbideetan:

- Interfazea: https://github.com/Elhuyar/BehaguneUI
- Datuen bilketa egiteko modulua:https://github.com/Elhuyar/MSM
- EliXa Polaritate modulua: https://github.com/Elhuyar/EliXa

Softwarea publiko egitearen helburua ez da soilik edonoren esku uztea, baizik eta etorkizunera begira, software hori hobetu eta erabiliko duen komunitate bat sortzea.

Software libreak zenbait abantaila sozial dakartza berekin, batez ez, bere erabilera eta banaketari dagokionean. Horrela, software libreak hurrengo lau askatasunak ematen dizkio erabiltzaileari:

- Programa nahi duen eran exekutatzeko askatasuna.

- Programaren iturburu kodea aztertu eta nahi duen eran aldatzeko askatasuna.
- Programaren kopia zehatzak banatzeko eskubidea.
- Aldatutako bertsioen kopiak banatzeko askatasuna.

Lau askatasun horiek oinarritzat hartuz erabiltzaileak norberaren konputazioaren kontrola bereganatu dezakete. Programaren banaketa eta erabilera ahalbidetzen duen sistema soziala bidezkoa da, eta, softwarea, banaketa eta erabilerari dagokionean, etikoa da.

4 Etorkizunerako Ondarea

"D2016 Behagunea" proiektuan garatutako software oro lizentzia libre baten bidez eskuragarri dagoela aipatu dugu. Proiektuaren baitan sortutako den aplikazioa edonork jarri dezake martxan, dela enpresa, erakunde, edo gizabanakoa.

Behatokiaren aplikazioaren oinarrizko osagaiak NLP moduluak dira, testuaren analisi linguistiko automatikoak gauzatzen dutenak, alegia. Modulu horiek egun eskuragarri daude lizentzia librean, eta beren garapena etengabe ari da gauzatzen. Behagunean tresnak D2016 domeinura egokitzeko egindako lan guztia komunitatearen esku dago jada.

Elhuyarren asmoa da, proiektu honetan garatzen den teknologia beste eszenatoki batzuetara zabaltzea. Zentzu horretan, kulturarekin lotutako ekintzak dira aplikazio eremu nagusietako bat, adibidez, Zinemaldia, Behobia-Donostia lasterketa, edo Tabakalera proiektuaren inguruko jarrera. Orokorrean, herritarren iritziak bideratzeko tresna izango da Behatokia, erakundeek ere herritarrengana hurbiltzeko baliatu dezaketena.

Hizkuntza-teknologien ikuspuntutik, Euskararen garapen digital eta teknologikorako urrats berri bat eman da, eta euskararako analizatzaile linguistikoak jarri ditu euskal gizartearen eskura.

5 Proiektuaren ebaluazioa

Memoria hau idazteko garaian, “DSS2016 Behagunea” ez da oraindik denbora nahikoa igaro garatutako soluzioa publiko denetik, eta beraz hemen aurkezten den ebaluazioa mugatua da. Gure ebaluazioa 3 ardatz nagusiren inguruan burutuko da: garatutako softwarea, Donostia 2016 proiektuaren martxa neurtzeko izan duen erabilgarritasuna, eta herritarren parte-hartze dinamikak bultzatzeko garaian izan duen eragina.

Garatutako softwareari dagokionez, sailkatzaileen ebaluazioa txosten honen bigarren atalean ematen da. Garatutako tresnen inguruko dokumentazioa amaitzen ari gara memoria hau idazteko unean, baina etiketatzaile linguistikoen inguruko dokumentazioa bakoitzari dagokion jatorrizko helbideetan aurki daiteke. Garatutako aplikazioa instalatzeko eta tresna guztien integraziorako dokumentazioa memoria hau osatzearekin batera argitaratuko da.

Bestalde, goiz da software garatzaileen komunitatean garatutako tresnek izan duten eragina neurtzeko. Ixa-pipes tresna multzoak hainbat jarraitzaile dituela konfirmatu dugu eta mundu osoko proiektuetan erabiltzen ari dela ere bai. Zentzu horretan, uztailan egingo den Codefest udako laborategian “Behagunea”-k parte-hartuko du proiektu gisa, eta dagoeneko badakigu interesa sortu duela proiektuak. Proiektua gizarteratzeko eta garatutako tresnak berrerabiltzeko bidean urratsa garrantzitsua dugu hau.

Azkenik Behaguneak DSS2016 proiektuaren ibilbidea aztertzeko tresna gisa ere diseinatu zen. Zentzu horretan, martxan daraman 4 hilabeteetan informazio baliotsua jaso du, proiektu ezberdinek une ezberdinetan izan duten oihartzuna neurtzen ari delarik. Behagunea publiko egin ez bada

ere, mundu digitalean sortzen diren iritziak monitorizatzeko duen ahalmena dagoeneko datu baliagarriak erauzteko balio duela erakusten ari da. Kuantitatiboki, 4 hilabeteetan 17.740 aipamen jaso ditu Behaguneak, 3.891 erabiltzaile ezberdinengandik (sare sozialetako atsegite eta bertxioak kontuan izan gabe). Epe honetan aipamen ezberdin gehien izan dituzten proiektuak azaltzen dira 4 taulan.

Proiektua	Aipamenak
Donostia_2016	11804
Euro-Dialogues	753
Emusik	735
Stop_War_Festibala	485
Gaur_Konstelazioak	441
Hirikilabs	379
Olatu_Talka	304
Elkarrizketak	267
Irekiera	190
2016_Bidea	157
Energia_Olatuak	125
Europa_Transit	115
Time_Machine_Soup	115
Labore	111
Rekording	107
Tosta	101
Hibrilaldiak	98
Tira_Birak	98
WhatsART	97

Taula 4. Oihartzun handiena lortu duten proiektuak, aipamen kopurua aipamen ezberdinei dagokie, sare sozialetako atsegite eta bertxioak kontuan izan gabe.

Hizkuntzei dagokienez, jasotako aipamenen %50 gaztelaniaz da, %37a euskaraz, %10 ingelesez eta gainontzeko %3a frantsesez.