IEB 2011, Donostiako Teknologi Elkartegian, 2011ko maiatzaren 12an

Barack Obamaren edo beste norbaiten tweet-etatik mamia ateraz

Igor Ruiz-Agundez, Izaskun Canga-Sanchez eta Marco Guidi

Deustuko Unibertsitatea, Univ. di Roma, Univ. del Salento

Dakizuenez baliabide sozialek orokorrean, eta gizarte-sareek bereziki, hedadura ikaragarria izan dute azken urteotan.

Gizarte-sareak datuz beterik daude:

Guk sareraturiko mezuak, argazkiak, dokumentuak, estekak, lagunak, gustuak, harremanak, fitxategiak eta abar.

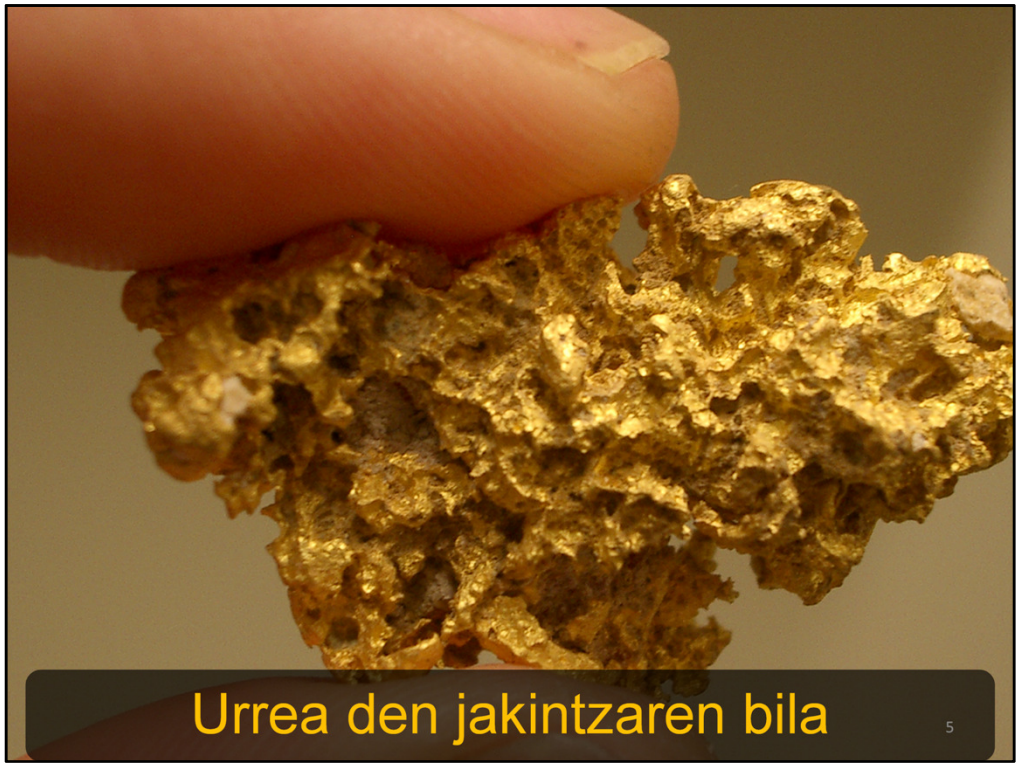
Datu horiei guztiei datu-meatzaritzako teknikak aplikatu ahal zaizkie.

Datu-meatzaritza (ingelesez, Data mining (DM)), datu kopuru handiak oinarri izanik, baliozko ezagutza eta informazioa eskuratzean datza.

Algoritmo egokiak erabiliz, datuen patroi edo eredu bat lortzea du helburu.

Datu-meatzaritza beste esparru askotan erabiltzen da:

Web bilaketetan, linguistikan, ekonomian, biologian, psikologian eta soziologian erabiltzen da batik bat.



Datu-meatzaritzak guretzat urrearen balioa duen jakintza eskuratzen laguntzen digu.

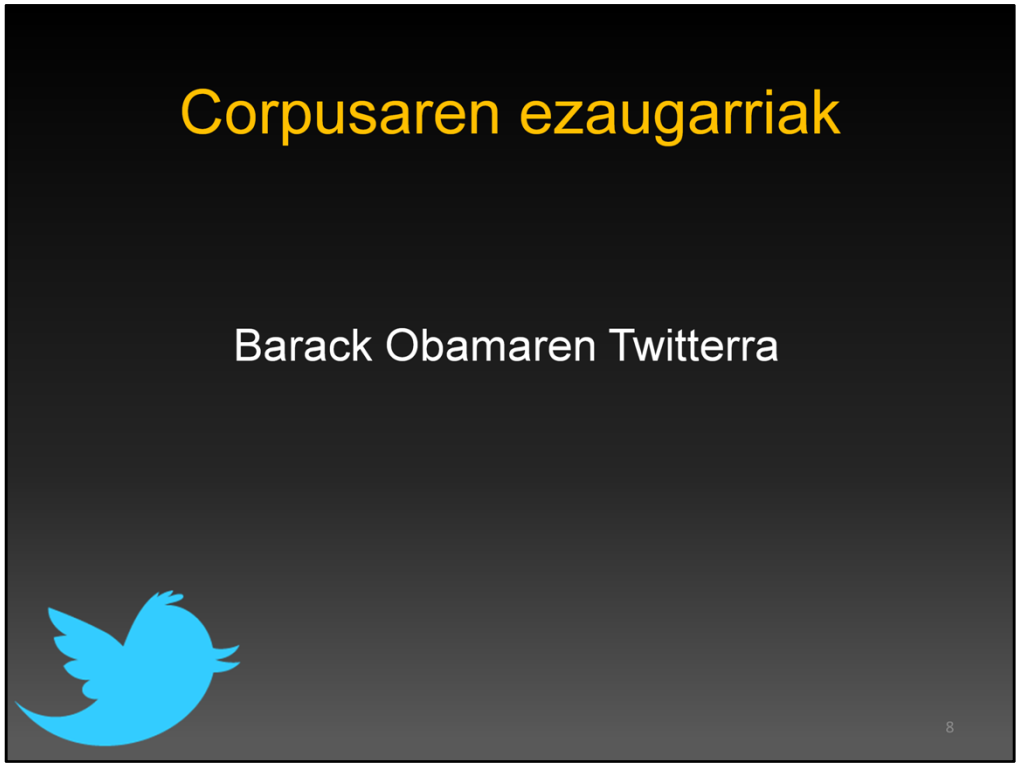
Horri berari esker, ditugun datuetatik jakintza eskuratuko dugu.

Gizarte-sareen kasuan, guretzat ezezagunak ziren erlazioak, esanahiak eta konnotazioak eskuratzen eta argitzen lagunduko digu.

Gizarte-sareak eta datu-meatzaritza batera jarri nahi genituen.

Horretarako, ondoren aurkeztuko dugun esperimentua egin dugu.

Aurkezpenik behar ez duen gizon honen Twitter gizarte-sareko mezuak erabili ditugu.



Egindako esperimentuaren corpus edo testu-datuak Barack Obamaren Twitterreko post, mezu edo tweetez osaturik dago.

Obama Estatu Batuetako senatari lanean hasi zenetik, hau da, 2007ko apirilaren 29tik, eta 2010eko irailaren 6 arte bildu dira datuak.

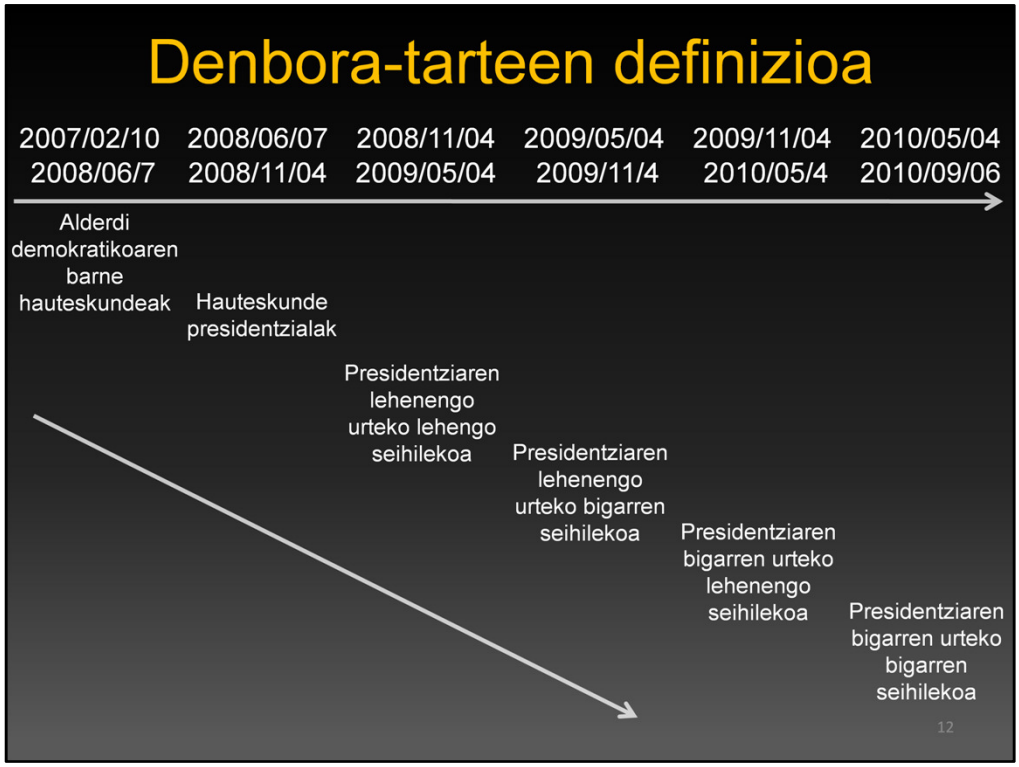
Guztira, 895 tweet, 16.000 karaktere baino gehiago eta 3.000 unitate lexikal inguruko corpusa lortu dugu.

Baina, nolakoa da tweet bat barrutik?
Zein ezaugarri ditu?

Tweet bakoitzak hainbat atributu ditu:

Jatorria: mezua nondik igortzen den zehazten du (ordenagailua, mugikorra, programa automatikoa, etab.).
Geo-lokalizazio mota eta geo-koordenatuak: mezua nondik igortzen den zehazten du.
Gogokoak: mezua “gogokoa” den edo norbaitek interesekotzat markatu duen zehazten du.
Identifikadoreak tweet bakoitza bakarra dela adierazten du.
Beste erabiltzaileekiko eta mezuekiko duen erlazioa zehazten du.
Mezuaren testua: partekaturiko 140 karatetako mezua.
Sortze ordua eta data.
Denbora absolutua: sortze denbora modu berezian kodifikatzen du.

Gure esperimentuan, tweeten testuaren edukia eta sorrera-data eta ordua bakarrik hartu ditugu kontuan.



Zehatzago esanda, mezu bakoitzak sorrera-momentu bat duenez, datak denbora-tarte jakinetan bildu ditugu, datuei esanahi zabalagoa emateko. Taula honek aipatutako denbora-tarteei buruzko azalpena ematen du.

Alderdi demokratikoaren barne hauteskundeak
Hauteskunde presidentzialak
Presidentziaren lehenengo urteko lehengo seihilekoa
Presidentziaren lehenengo urteko bigarren seihilekoa
Presidentziaren bigarren urteko lehenengo seihilekoa
Presidentziaren bigarren urteko bigarren seihilekoa

Datu-meatzaritzan erabilitako tresna

Datu hauei testu-analisirako T-Lab software-tresna aplikatu diegu.

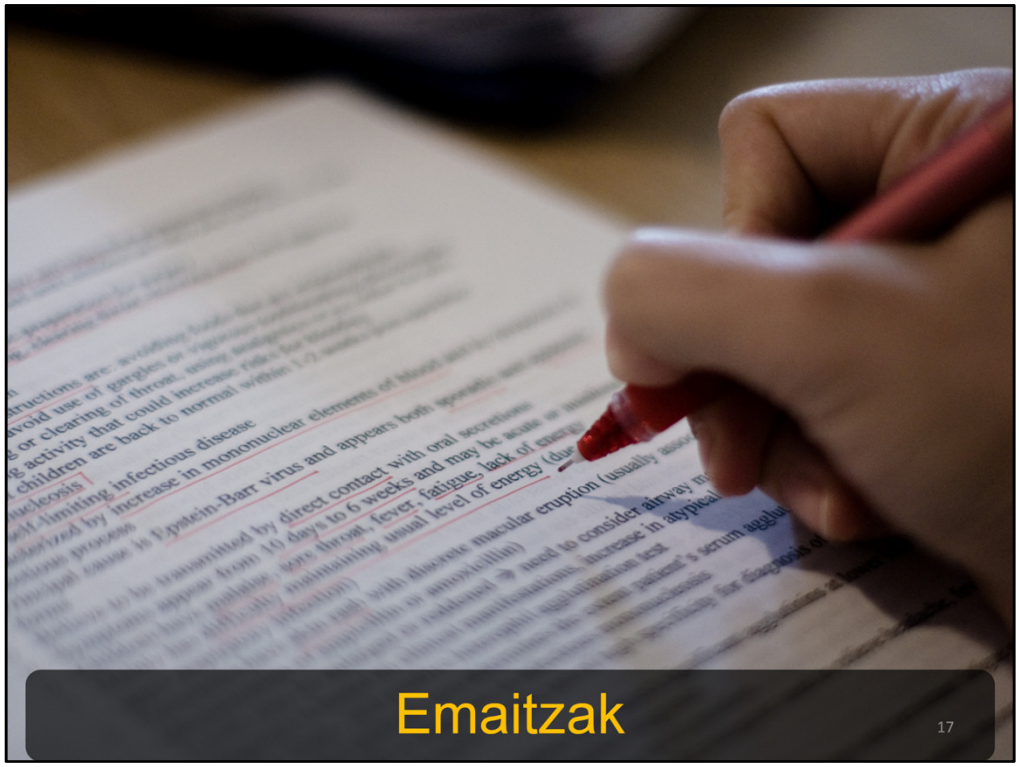
Software horrek testu-edukien analisia eta testu-meatzaritza ahalbidetzen ditu, eta gure esperimenturako beharrak bete.

Guk erabilitako tresnaren gaitasunak edozein eratako dokumentuetan aplikatu daitezke (diskurtsoak, egunkarietako artikuluak, inkestak, liburuak, etab.).

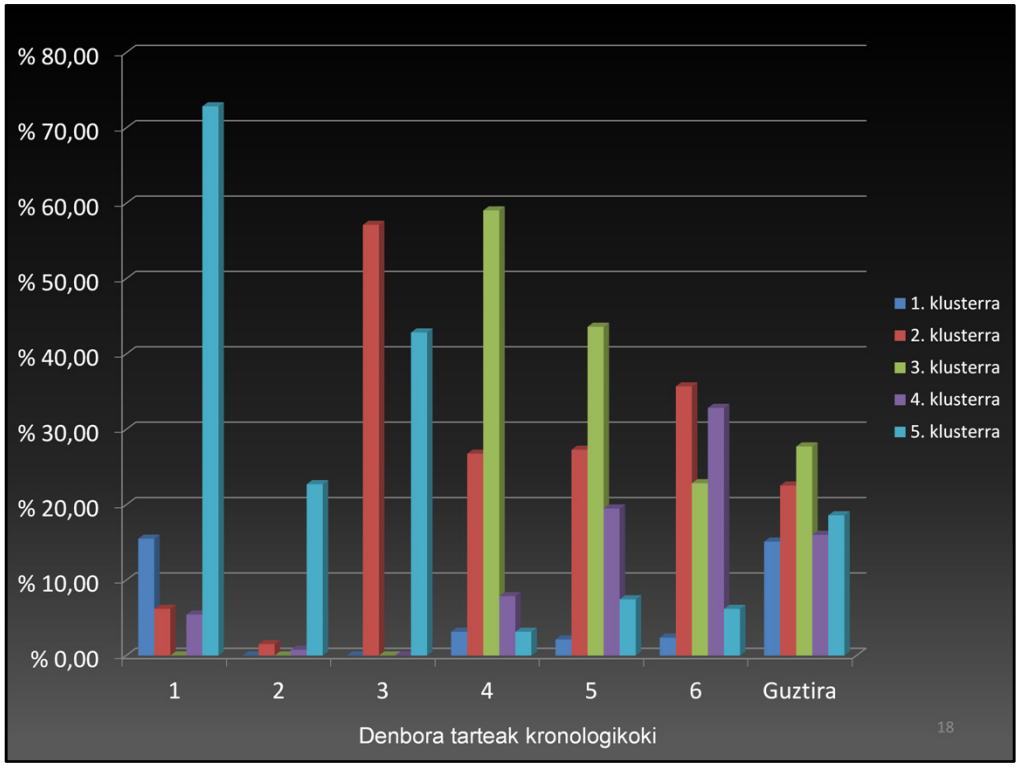
Orokorrean, testuz osaturiko edozein elementuri aplikatu dakioke.

Guk dakigunetik, inork ez ditu, orain arte, gizarte-sareetako testuak modu horretan ikertu.

Egin daitezkeen analisi mota guztien artean, Tweeter mezuen oinarrizko testuinguruak aztertzea erabaki dugu.



Tweetak eskuratu, prestatu, moldatu eta txukundu ostean aipaturiko gaien analisia oinarrizko testuinguruen bidez burutu dugu, eta ondoren aurkezten diren emaitzak lortu ditugu.



Analisi horrek bilduriko tweetak multzokatuko ditu esanahi berria emanez. Goiko taulak kluster guztien distribuzioa erakusten du.

Kluster bakoitza aztertuz gero, ezaugarri eta eduki propioak dituela antzeman daiteke eta bakoitzean hitz eta esaldi esanguratsu propioak daudela.

Informazio horrekin (Twitterreko mezuetatik eta analizaturiko denbora-tarte bakoitzean garrantzi gehien duen klusterretik), esanahi-sortze prozesu bat egin dugu eta Barack Obamak buruturiko politikarekiko korrespondentzia bat aurkitu dugu.

Modu honetan, denbora-tarte bakoitzak pareko egoera politiko bat izan du:

Aldaketa sustatu

Hauteskunde presidentzialen denbora-tartean emaniko kluster honek aldaketa sustatzen duela interpretatu dugu.

Obamaren proposamen politikoa aldaketarekin identifikatu da hasieratik.

"Change We Need", "Vote for Change" (Aldaketa behar dugu, bozkatu aldaketaren alde) lelopean buruturiko kanpaina izan du.

Twitterren, hautesleak ekitaldi politikoak Interneten bidez jarraitzera gonbidatzen ditu.


live
watch
rally
earlythe
it
clinton
change
for
event
joe
meet
http://my.barackobama.com
reno
energy
we
conference
town
livestream
stream
discussion
economic
hold
need
hall
vote
20

Amerikarren lehentasunak

Presidentziaren lehenengo urteko lehengo seihilekoan emaniko kluster honek Amerikarren lehentasunak islatzen dituela interpretatu dugu.

Amerikarren lehentasunak islatzen dituzten mezuak biltzen ditu talde honek: etorkizuna, Amerikar Nazioa, konpromisoa, lana, borroka, energia garbiak, eskubideak, etab.


lose
pass
future
energy
rights
recovery
freedom
people
american
pay
commitment
historic
create
president
combat
mission
cleangulf
obama
work
coast
22

Osasun erreforma

Presidentziaren lehengo urteko bigarren seihilekoan eta bigarren urteko lehenengo seihilekoan emaniko kluster honek osasun erreforma islatzen duela interpretatu dugu.

Talde honek gai bakar bat aurkezten du.

Osasun erreformaren inguruan emaniko pausuak komentatzen dira, eta oso gai eztabaidatua dela ikus daiteke.


america
final
law
organizing
reform
wall
yesterday
congress
organize
street
important
spread
protection
insurance
member
hear
health
senate
http://bit.ly
for
video
consumer
of
a
call
24

Gizarte gaiak

Presidentziaren bigarren urteko bigarren seihilekoan emaniko kluster honek gizarte gaiak islatzen dituela interpretatu dugu.

Talde honek hautesleei Obamaren estrategia politikoaren oinarriak jakinarazteko mezuak biltzen ditu.

Gizarte gaiak dira aipatuenak: ekonomia, ingurumena eta segurtasuna.


duncan
speak
best
nuclear
house
job
economy
forward
question
listen
http://wh.gov
innovation
national
move
live
education
answer
white
tune
watch
arne
26

Parte hartzea sustatzen

Alderdi demokratikoaren barne hauteskundeen denbora-tartean bilduriko mezuek osatzen duten klusterreko edukiak jendeak ekitaldietan parte hartzea sustatzen du.

Azken gai honek hauteskunde garaian parte hartzea bultzatzeko mezuak biltzen ditu.

Obamak aldaketarekin duen konpromisoa azpimarratzen da eta mitin, hitzaldi eta elkarrizketak jarraitzeko deia ere egiten da.


info
msnbc
look thought
presidential tonight change
http://tinyurl.com text
debate
poll
http://barackobama.com
today city iowa campaign
full head visit
ask
28

Twitter gizarte-sarearen mezuen analisirako metodologia bat aurkeztu dugu.

Honi esker, Barack Obamaren tweetetatik esanahi-sortze prozesu bat egiteko gai izan gara.

Emaitzei esker, Obamaren mezuak bost talde ezberdinetan bildu ditugu.

Horrela, denbora-tarte bakoitzari korrespondentzia politiko bat aurkitu diogu, klusterizazioaren emaitzei esker.

Are gehiago, tweeten edukiak gure interpretazioarekin bat egiten duela ikus daiteke.

Gizarte-sareetatik benetako bizitza islatzen duen informazio esanguratsua ateratzeko aukera dagoela ikusi dugu.

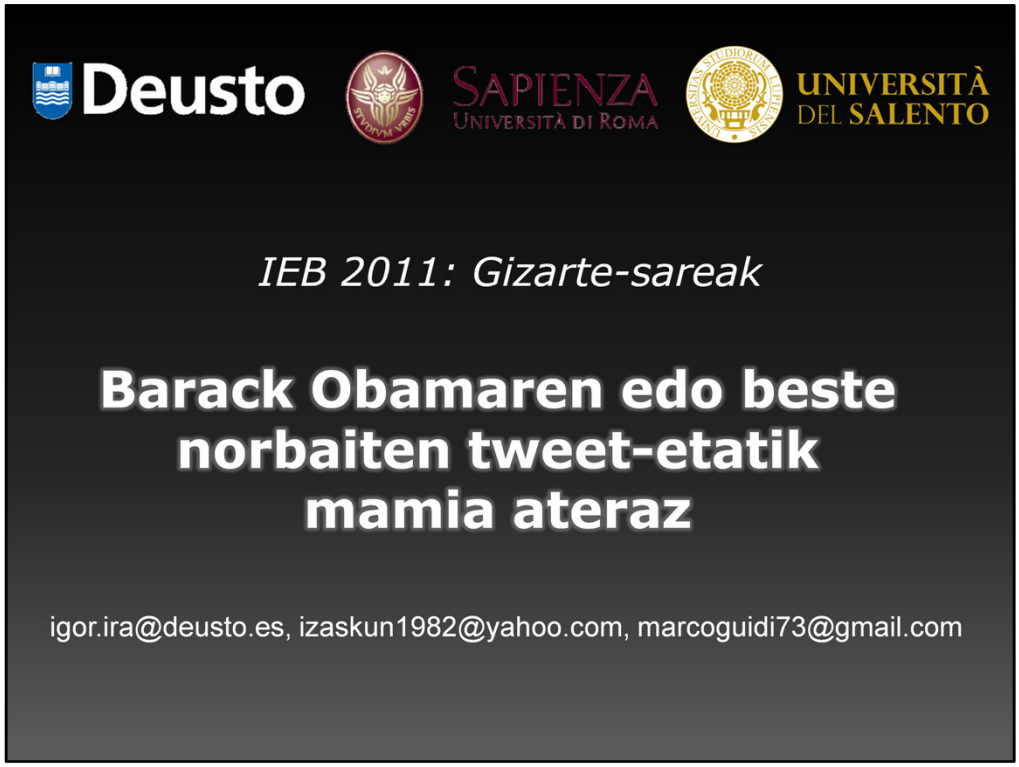
Etorkizun hurbilean, aurkezturiko lana beste Twitter erabiltzaile batzuei aplikatzea pentsatu dugu.

Aurkezturiko lana tweet mezu baten partekatzea erabakitzen baduzue zera esan dezakezue:

"Gizarte-sareetatik jakinduria eskura dezakegu #ieb11".

Mezuak 140 karaktere baino gutxiago dituela gogora ezazue.

Eskerrik asko.



IEB 2011, Donostiako Teknologi Elkartegian, 2011ko maiatzaren 12an

Barack Obamaren edo beste norbaiten tweet-etatik mamia ateraz

Igor Ruiz-Agundez, Izaskun Canga-Sanchez eta Marco Guidi

Deustuko Unibertsitatea, Univ. di Roma, Univ. del Salento