Entrega: 9/10/2023 - Pròrroga: 18/10/2023
Noms: Mariona Farré Tapias Marc Pérez Guerrero

Activitat 2: Rein

Introducció:

Per poder executar els programes que utilitzem a continuació, és important que tinguem el programa ElasticSearch en execució:

```
##descarregar el programa en l'ordinador
https://www.elastic.co/es/downloads/elasticsearch
##comanda per localitzar el programa
whereis elasticsearch
##executar-ho en funció on estigui: cas ordinadors d'universitat:
 /usr/share/elasticsearch/bin/elasticsearch
```

Si tot està correctament creat, a l'executar la següent instrucció no hauria de donar cap error:

```
%run elastic_test.py
```

Recàrrega de l'índex:

Analitzar l'efecte sobre un índex si aquest se li fan canvis sobre la tokenització dels termes i filtres addicionals per els termes per normalitzar i filtrar tots aquells que no siguin útils per la búsqueda o el treball sobre un índex d'un corpus.
Donat un nou fitxer de laboratori *IndexFilesPreprocess.py* li podem canviar diferents paràmetres amb els arguments de *--token* i *--filter.*

Tokenització:

Al fitxer *IndexFilesPreprocess.py* si li posem en l'argument de token aquest té diferents opcions per diferents funcions, amb això podem observar els diferents canvis entre els índex creats.
Procés de creació la tokenització:
Executar el fitxer IndexFilesPreprocess.py amb el filtre de whitespace i el path (camí) de on es trobi el corpus, en el nostre cas la col·lecció de novel·les.

Executar: python3 IndexFilesPreprocess.py --index lab2_ws --path REIN/LAB2/novels --token whitespace

```
_ws  --path                                    python3 IndexFilesPreprocess.py --index lab2
                                               0/REIN/LAB2/novels --token whitespace
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarnin
g: the default value for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in v
ersion 8; specify '?wait_for_active_shards=index-setting' to adopt the future default behaviour, or '?wait_
for_active_shards=0' to preserve today's behaviour
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarnin
g: [types removal] Using include_type_name in put mapping requests is deprecated. The parameter will be rem
oved in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_ws': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preferenc
e': 'data_content'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_ws', 'creation_date': '1696945349268
', 'analysis': {'analyzer': {'default': {'filter': ['lowercase'], 'type': 'custom', 'tokenizer': 'whitespac
e'}}}, 'number_of_replicas': '1', 'uuid': 'sQLCItbrS6CvxfPWqvv7aw', 'version': {'created': '7171299'}}}}}
Indexing
```

Després per poder veure el nombre de tokens creats executar el fitxer *CountWords.py* en l'index creat i guardar-ho en un fitxer tipus .csv:
Executar: time p thon3 CountWords.p --index lab2_ws > lab2_ws.csv

```
/REIN/LAB2$ time python3 CountWords.py --index lab2_ws_f > lab2_ws_f.csv

real    0m9,038s
user    0m4,403s
sys     0m0,317s
```

Fer el mateix procés per tots els tipus de tokenització, fins aconseguir tots els fitxers .csv que corresponguin a cada tipus:

```
REIN/LAB2$ python3 IndexFilesPreprocess.py --index lab2_cl  --path /hom
/LAB2/novels --token classic
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: the default v
alue for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in version 8; specify '?wait_for_act
ive_shards=index-setting' to adopt the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behavio
ur
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: [types remova
l] Using include_type_name in put mapping requests is deprecated. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_cl': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preference': 'data_conten
t'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_cl', 'creation_date': '1696945977070', 'analysis': {'analyzer': {'de
fault': {'filter': ['lowercase'], 'type': 'custom', 'tokenizer': 'classic'}}}, 'number_of_replicas': '1', 'uuid': '1cT_wPQd
TBiA6RzTjix4uw', 'version': {'created': '7171299'}}}}}
Indexing ...
REIN/LAB2$ time python3 CountWords.py --index lab2_cl > lab2_cl.csv

real    0m8,680s
user    0m5,858s
sys     0m0,333s
```

```
REIN/LAB2$ python3 IndexFilesPreprocess.py --index lab2_st  --path /hom
LAB2/novels --token standard
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: the default v
alue for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in version 8; specify '?wait_for_act
ive_shards=index-setting' to adopt the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behavio
ur
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: [types remova
l] Using include_type_name in put mapping requests is deprecated. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_st': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preference': 'data_conten
t'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_st', 'creation_date': '1696946092635', 'analysis': {'analyzer': {'de
fault': {'filter': ['lowercase'], 'type': 'custom', 'tokenizer': 'standard'}}}, 'number_of_replicas': '1', 'uuid': 'Zd7lssq
kREqajU0z2ziQ4A', 'version': {'created': '7171299'}}}}}
Indexing ...
REIN/LAB2$ time python3 CountWords.py --index lab2_st > lab2_st.csv

real    0m7,860s
user    0m4,872s
sys     0m0,279s
```

```
REIN/LAB2$ python3 IndexFilesPreprocess.py --index lab2_lt  --path /hom
/LAB2/novels --token letter
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: the default v
alue for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in version 8; specify '?wait_for_act
ive_shards=index-setting' to adopt the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behavio
ur
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: [types remova
l] Using include_type_name in put mapping requests is deprecated. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_lt': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preference': 'data_conten
t'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_lt', 'creation_date': '1696946288887', 'analysis': {'analyzer': {'de
fault': {'filter': ['lowercase'], 'type': 'custom', 'tokenizer': 'letter'}}}, 'number_of_replicas': '1', 'uuid': 'EgsYN0UGT
VmQHvThCZTALw', 'version': {'created': '7171299'}}}}}
Indexing ...
REIN/LAB2$ time python3 CountWords.py --index lab2_lt > lab2_lt.csv

real    0m7,660s
user    0m4,657s
sys     0m0,372s
```

Els fitxers creats haurien de ser els següents:

Es poden veure els índex creats també entrant a: http://localhost:9200/_cat/indices

```
localhost:9200/_cat/indices

green  open .geoip_databases TPPdUBXSSsKgIjef9asyQQ 1 0    42 0 40.6mb 40.6mb
yellow open lab2_ws          CUm-Sou_SP-B-g0yzKIOXA 1 1 20089 0 38.7mb 38.7mb
yellow open lab2_lt          pVhZcSKfTMebRtQYtSXguQ 1 1 20089 0   35mb   35mb
yellow open lab2_cl          P7RlKQCfReK3r2_X60ZpnQ 1 1 20089 0 35.7mb 35.7mb
yellow open lab2_st          _O3joEf6Sl2puA95v_7QuQ 1 1 20089 0 35.5mb 35.5mb
```

El resultat dels .csv son els següents:

- **Whitespace**:

El tokenització *whitespace* divideix el text en termes cada vegada que detecti un espai en blanc.

	A
1	1 _socrates 1, "_when 1, "accident,
2	1 accurate 1, "analogical
3	1 angels
4	1 answer": 1, "aristotle 1, "association 1, "beggar,
5	1 big
6	1 black-bird,
7	1 book,
8	1 c 1, "canon".

94562	3	welch
94563	3	westwood
94564	3	widely-different
94565	3	widely-diffused
94566	3	wingless
94567	3	youatt
94568	4	alexander 4, "all
94569	4	concerning 4, "hypothetical 4, "law 4, "p 4, "whiteness
94570	4	(_a_)
94571	4	(_c_)
94572	4	(another
94573	4	(whether
94574	4	110.
94575	4	131
94576	4	1831

```
24588, with
24860, it
26406, he
27474, his
32123, i
33298, was
34288, that
56447, in
63102, a
82329, to
97314, and
115552, of
203231, the
-----------------------------------------------
167063 Distinct words, 3099105 Word occurrences
```

- **Classic**:

El tokenització *classic* divideix el text segons la gramatització del llenguatge anglès.

	A	B
1	1	1/20
2	1	100-115
3	1	
4	1	
5	1	1011_b
6	1	111-1/9
7	1	111-à
8	1	111th
9	1	112th
10	1	121-202
11	1	
12	1	1276-7
13	1	1452-1519
14	1	1473-1543
15	1	1508-1588

44873	10	comas
44874	10	aberrant
44875	10	anthers
44876	10	grafting
44877	10	interbreeding
44878	10	naturalisation
44879	10	polymorphic
44880	10	tarsi
44881	11	109
44882	11	158
44883	11	1607
44884	11	1832
44885	11	1833
44886	11	1913
44887	11	199
44888	11	295
44889	11	297
44890	11	acceptation

59559	28034	he
59560	31712	it
59561	34648	was
59562	37376	that
59563	37934	i
59564	57988	in
59565	65288	a
59566	84215	to
59567	101593	and
59568	116802	of
59569	206706	the
59570	----------	----------
59571	59569 Distinct words	3144033 Word occurrences

- **Standard**:

El tokenització *standard* divideix el text en termes segons els límits de paraules, definit per l'algoritme Unicode Text Segmentation, elimina els signes de puntuació.

	A	B
1	1	aar
2	1	absoluta
3	1	absolutus
4	1	academe
5	1	accentus
6	1	accidentia
7	1	accidentis
8	1	accipio
9	1	acoustic
10	1	acropolis

45173	8	hilaire
45174	8	intertropical
45175	8	lobelia
45176	8	natura
45177	8	reversions
45178	8	saltum
45179	8	wingless
45180	9	347
45181	9	78
45182	9	_a_
45183	9	_ad
45184	9	_c_
45185	9	_e_
45186	9	_here_

61812	24993	with
61813	25167	as
61814	27824	his
61815	27976	he
61816	31682	it
61817	34576	was
61818	37280	that
61819	37791	i
61820	57955	in
61821	65238	a
61822	84200	to
61823	101586	and
61824	116786	of
61825	206546	the
61826	----------	----------
61827	61825 Distinct words	3147864 Word occurrences

- **Letter**:

El tokenització *letter* divideix el text en termes segons si el caràcter que es troba és o no és una lletra.

	A	B
1	1	1006
2	1	1008
3	1	1011_b_
4	1	111th
5	1	112th
6	1	1265
7	1	1276
8	1	1452
9	1	1473
10	1	1508
11	1	1543
12	1	1894_
13	1	1899_
14	1	1904_
15	1	1909_
16	1	1912_

46786	28	newty
46787	28	teton
46788	28	walkullas
46789	28	larvae
46790	29	acknowledgment
46791	29	ad
46792	29	agility
46793	29	akin
46794	29	annually
46795	29	boiled
46796	29	boiling

54445	28666	he
54446	32740	it
54447	34649	was
54448	38080	that
54449	39900	i
54450	57991	in
54451	65487	a
54452	84279	to
54453	101593	and
54454	116803	of
54455	206706	the
54456	----------	----------
54457	54455 Distinct words	3160329 Word occurrences

Els resultats són:

tokenització	*Diferents paraules*	*Nombre de paraules trobades*
Whitespace	167063 Distinct words	3099105 Word occurrences
Classic	59569 Distinct words	3144033 Word occurrences
Standard	61825 Distinct words	3147864 Word occurrences
Letter	54455 Distinct words	3160329 Word occurrences

Com podem observar en la taula anterior, tots els tipus de tokenització es troben casi el mateix número de paraules al voltant d'unes 3100000, el que varia és el nombre de paraules diferents, que dependrà de com d'estricte sigui el tokenizador, en aquest cas el que distingeix menys paraules és el *letter* i el que més el *whitespace*.
Sent la paraula amb més freqüència "the" seguida de diferents tipus pronoms i articles de la llengua anglesa.

Filtres:

Al fitxer *IndexFilesPreprocess.py* si li posem en l'argument de *filter* aquest té diferents opcions per diferents funcions de segmentació com lowercase, asciifolding, stop etc.
Cada filtre té diferents funcionalitats, nosaltres hem trobat que els filtres més útils per el nostre corpus de novel·les en llengua anglesa:

- **lowercase:** passa totes les lletres de majúscules a minuscules.
- **asciifolding:** converteix lletres,números i símbols en els seus ASCII equivalents si existeixen, per exemple amb els accents "à" passa a ser "a".
- **stop:** elimina les *stopwords* dels tokens, paraules que no tenen cap valor per fer un índex , per exemple: "a", "in","not","the","with" etc.
- **kstem:** algorisme de stemming per la llengua anglesa, combina un algorisme de stemming i un diccionari intern.

Procés de creació la tokenització amb filtres:
Executar el fitxer IndexFilesPreprocess.py amb el filtre de whitespace i el path (camí) de on es trobi el corpus, en el nostre cas la col·lecció de novel·les.

Executar: python3 IndexFilesPreprocess.py --index lab2_ws_f --path /REIN/LAB2/novels --token whitespace --filter lowercase asciifoldin sto kstem

```
REIN/LAB2$ python3 IndexFilesPreprocess.py --index lab2_ws_f  --path /home
AB2/novels --token whitespace --filter lowercase asciifolding stop kstem
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: the default valu
e for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in version 8; specify '?wait_for_active_sh
ards=index-setting' to adopt the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behaviour
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: [types removal]
Using include_type_name in put mapping requests is deprecated. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_ws_f': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preference': 'data_content
'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_ws_f', 'creation_date': '1696951404699', 'analysis': {'analyzer': {'defa
ult': {'filter': ['lowercase', 'asciifolding', 'stop', 'kstem'], 'type': 'custom', 'tokenizer': 'whitespace'}}}, 'number_of_re
plicas': '1', 'uuid': 'o3OcH2wQTgW8noNx3DmD6w', 'version': {'created': '7171299'}}}}}
Indexing ...
```

Després per poder veure el nombre de tokens creats executar el fitxer *CountWords.py* en l'index creat i guardar-ho en un fitxer tipus .csv:

Executar: time python3 CountWords.py --index lab2_ws > lab2_ws.csv

```
REIN/LAB2$ time python3 CountWords.py --index lab2_ws_f > lab2_ws_f.csv

real    0m6,931s
user    0m4,405s
sys     0m0,285s
```

Fer el mateix procés per tots els tipus de tokenització, fins aconseguir tots els fitxers .csv que corres on uin a cada ti us:

```
REIN/LAB2$ python3 IndexFilesPreprocess.py --index lab2_cl_f  --path /hom
LAB2/novels --token classic --filter lowercase asciifolding stop kstem
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: the default val
ue for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in version 8; specify '?wait_for_active_
shards=index-setting' to adopt the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behaviour
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: [types removal]
 Using include_type_name in put mapping requests is deprecated. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_cl_f': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preference': 'data_conten
t'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_cl_f', 'creation_date': '1696952242174', 'analysis': {'analyzer': {'de
fault': {'filter': ['lowercase', 'asciifolding', 'stop', 'kstem'], 'type': 'custom', 'tokenizer': 'classic'}}}, 'number_of_re
plicas': '1', 'uuid': 'he3ielEZQiueOqRNG2jW5A', 'version': {'created': '7171299'}}}}}
Indexing ...
REIN/LAB2$ time python3 CountWords.py --index lab2_cl_f > lab2_cl_f.csv

real    0m6,390s
user    0m3,623s
sys     0m0,300s
```

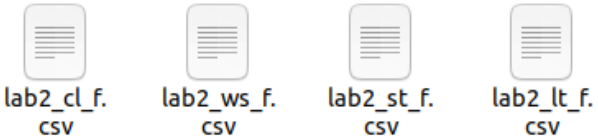
```
REIN/LAB2$ time python3 CountWords.py --index lab2_st_f > lab2_st_f.csv
Traceback (most recent call last):
  File "CountWords.py", line 37, in <module>
    for s in sc:
  File "/usr/local/lib/python3.8/dist-packages/elasticsearch/helpers/actions.py", line 494, in scan
    resp = client.search(
  File "/usr/local/lib/python3.8/dist-packages/elasticsearch/client/utils.py", line 139, in _wrapped
    return func(*args, params=params, headers=headers, **kwargs)
  File "/usr/local/lib/python3.8/dist-packages/elasticsearch/client/__init__.py", line 1479, in search
    return self.transport.perform_request(
  File "/usr/local/lib/python3.8/dist-packages/elasticsearch/transport.py", line 345, in perform_request
    status, headers_response, data = connection.perform_request(
  File "/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/http_urllib3.py", line 256, in perform_request
    self._raise_error(response.status, raw_data)
  File "/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py", line 287, in _raise_error
    raise HTTP_EXCEPTIONS.get(status_code, TransportError)(
elasticsearch.exceptions.TransportError: TransportError(503, 'search_phase_execution_exception', '[aul-1605][127.0.0.1:9300][
indices:data/read/search[phase/query]]')

real    0m0,445s
user    0m0,535s
sys     0m0,135s
```

```
REIN/LAB2$ python3 IndexFilesPreprocess.py --index lab2_lt_f  --path /h
N/LAB2/novels --token letter --filter lowercase asciifolding stop kstem
Indexing 33 files
Reading files ...
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: the default v
alue for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in version 8; specify '?wait_for_act
ive_shards=index-setting' to adopt the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behavio
ur
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/usr/local/lib/python3.8/dist-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarning: [types remova
l] Using include_type_name in put mapping requests is deprecated. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'lab2_lt_f': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preference': 'data_cont
ent'}}}, 'number_of_shards': '1', 'provided_name': 'lab2_lt_f', 'creation_date': '1696953458558', 'analysis': {'analyzer':
{'default': {'filter': ['lowercase', 'asciifolding', 'stop', 'kstem'], 'type': 'custom', 'tokenizer': 'letter'}}}, 'number_
of_replicas': '1', 'uuid': 'ZHijiOWnScuG06IWvmAgiA', 'version': {'created': '7171299'}}}}}
Indexing ...
REIN/LAB2$ time python3 CountWords.py --index lab2_lt_f > lab2_lt_f.csv

real    0m5,599s
user    0m3,434s
sys     0m0,196s
```

Els fitxers creats haurien de ser els següents:



I en el local index sortiran els següents:

```
green  open .geoip_databases kzvD8F-QQ6O5c74jW0Q1qg 1 0 43 0 40.8mb 40.8mb
yellow open lab2_ws          sQLCItbrS6CvxfPWqvv7aw 1 1 33 0 19.2mb 19.2mb
yellow open lab2_st_f        Gb6_WZTRS3at6hNxw7Y8RQ 1 1 33 0 16.3mb 16.3mb
yellow open lab2_ws_f        o3OcH2wQTgW8noNx3DmD6w 1 1 33 0   18mb   18mb
yellow open lab2_cl_f        he3ieiEZQiueOqRNG2jW5A 1 1 33 0 16.3mb 16.3mb
yellow open lab2_lt          EgsYN0UGTVmQHvThCZTALw 1 1 33 0 17.8mb 17.8mb
yellow open lab2_cl          1cT_wPQdTBiA6RzTjix4uw 1 1 33 0 17.9mb 17.9mb
yellow open lab2_lt_f        x9uOt02dRmejtwGzBgkxMg 1 1  0 0   249b   249b
yellow open lab2_st          Zd7lssqkREqajU0z2ziQ4A 1 1 33 0 17.9mb 17.9mb
```

El resultat dels .csv son els següents:

- **Whitespace amb els filtres:**

	A	B
1		socrates 1, "_when 1, "accident,
2		accurate 1, "analogical
3	1	angels
4		answer": 1, "aristotle 1, "association 1, "beggar,
5	1	big
6	1	black-bird,
7	1	book,
8		c 1, "canon". 1, "cat,
9		cause". 1, "cause".--methods 1, "cavendish 1, "certainly,
10	1	circumstance,
11	1	class
12	1	class
13		classes 1, "cleverness,
14	1	colligation

```
8100, mr.
8194, so
8240, one
8506, been
8701, said
8868, she
10299, her
10879, were
11062, we
11216, all
12306, my
13787, you
13790, from
14521, which
16320, have
20472, had
26411, he
27474, his
32123, i
-----------------------------------------------
153187 Distinct words, 2082003 Word occurrences
```

- **Classic amb els filtres:**

	A	B
1	1	1/20
2	1	100-115
3	1	1006
4	1	1008
5	1	1011_b
6	1	111-1/9
7	1	111-a
8	1	111th
9	1	112th
10	1	121-202
11	1	1265
12	1	1276-7
13	1	1452-1519
14	1	1473-1543
15	1	1508-1588
16	1	1520-1603
17	1	1540-1603
18	1	19-23
19	1	23-4
20	1	3/1/7/9/31796

43078	10500	said
43079	10578	him
43080	11293	were
43081	11705	we
43082	11965	her
43083	12754	all
43084	12954	my
43085	14025	from
43086	15442	which
43087	16723	have
43088	17638	you
43089	20898	had
43090	27856	his
43091	28040	he
43092	37983	i
43093	---	---
43094	43092 Distinct words	2079573 Word occurrences

- **Standard amb els filtres:**

	A	B
1	1	1006
2	1	1008
3	1	1011_b
4	1	111th
5	1	112th
6	1	1265
7	1	1276
8	1	1452
9	1	1473
10	1	1508
11	1	1543
12	1	1894_
13	1	1899_
14	1	1904_
15	1	1909_
16	1	1912_
17	1	1913

45395	11286	were
45396	11684	we
45397	11904	her
45398	12719	all
45399	12922	my
45400	14012	from
45401	15440	which
45402	16712	have
45403	17540	you
45404	20869	had
45405	27824	his
45406	27982	he
45407	37838	i
45408	---	---
45409	45407 Distinct words	2084435 Word occurrences

- **Letter amb els filtres:**

	A	B
1	1	aar
2	1	absoluta
3	1	absolutu
4	1	academe
5	1	accentu
6	1	accidenti
7	1	accidentia
8	1	accipio
9	1	acoustic
10	1	acropolis
11	1	actio
12	1	addit
13	1	adioristo
14	1	adjectif
15	1	admitti
16	1	aff
17	1	affirmo
18	1	agnoia

37879	10579	him
37880	11294	were
37881	11934	we
37882	11969	her
37883	12765	all
37884	12954	my
37885	14025	from
37886	15443	which
37887	16723	have
37888	18040	s
37889	18465	you
37890	20899	had
37891	27861	his
37892	28672	he
37893	39949	i
37894	---	---
37895	37893 Distinct words	2093150 Word occurrences

Els resultats són:

tokenització	*Paraules diferents*	*Paraules diferents+filtres*	*Nombre de paraules trobades*	*Nombre de paraules +filtres*
Whitespace	167063 Distinct words	153187 Distinct words	3099105 Word occurrences	2082003 Word occurrences
Classic	59569 Distinct words	43092 Distinct words	3144033 Word occurrences	2079573 Word occurrences
Standard	61825 Distinct words	45407 Distinct words	3147864 Word occurrences	2084435 Word occurrences
Letter	54455 Distinct words	37893 Distinct words	3160329 Word occurrences	2093150 Word occurrences

Com podem observar en aquesta taula, és la comparació entre les diferents tokenitzacions i si s'han aplicat filtres (en color blau) o no. Com es pot veure, hi ha una considerable reducció de nombre de paraules de ≈ 3100000 a ≈ 2100000 aplicant els filtres, numeros que justifiquen que els filtres s'han aplicat correctament i han pogut treure unes 1000000 que no arriben els criteris *(lowercase, asciifolding, stop, kstem*) que hem decidit nosaltres que no volem tenir en l'índex. Amb les paraules diferents podem veure que han baixat considerablement de nombre al voltant d'unes 1000000 paraules, continuant sent el tokenizador que distingeix menys paraules el *letter* i el que més el *whitespace*.
Ara la paraula més freqüent passa de ser "the" a "i", una paraula que no es considera una stopword, al ser un pronom és una paraula que s'usa amb molta freqüència, les que tenen més son de pronoms "he","her","his","my".. , verbs com "have" o paraules molt usades com "from".

Càlcul dels pesos tf-idf i mesura del cosinus:

Per poder calcular els pesos tf-idf de dos documents dins del mateix corpus de documents, s'ha de calcular primerament els vectors tf-idf corresponent a dos documents i a continuació es pot calcular la seva similitud amb la mesura del cosinus entre els dos vectors.

Donat un fitxer de laboratori *TFIDFViewer.py* amb els paràmetres *--índex* se li passa el index del corpus donat, *--files* amb dos paràmetres dels dos camins on es troben els dos documents desitjats i l'argument *--print* per si es vol veure per pantalla el vector paraules de cada document, s'ha de completar per al executar aquest fitxer ens donés la similitud entre dos noveles.

Hem completat així les següents funcions del fitxer *TFIDFViewer.py* :
(Dins del fitxer esta comentat com funciona el codi especificament)

```python
def toTFIDF(client, index, file_id):
    # Obtenemos el vector de términos y el df (frecuencia de documentos) de los términos
    file_tf, file_df = document_term_vector(client, index, file_id)

    # Calculamos el número de documentos en el índice
    dcount = doc_count(client, index)

    # Calculamos el máximo valor de frecuencia en el documento
    max_freq = max([f for _, f in file_tf])

    # Inicializamos una lista para almacenar los pesos TF-IDF
    tfidfw = []

    for (t, w), (_, df) in zip(file_tf, file_df):
        # Calculamos el TF-IDF para cada término
        tf = w / max_freq  # TF (frecuencia de término)
        idf = np.log(dcount / df)  # IDF (frecuencia inversa de documento)
        tfidf = tf * idf  # TF-IDF

        # Agregamos el término y su peso TF-IDF a la lista
        tfidfw.append((t, tfidf))

    # Normalizamos el vector de peso TF-IDF usando NumPy
    tfidfw_array = np.array([w for _, w in tfidfw])
    magnitude = np.linalg.norm(tfidfw_array)
    normalized_tfidfw = [(t, w / magnitude) for t, w in tfidfw]

    return normalized_tfidfw


def normalize(tw):
    # Extraemos los pesos del vector en un array NumPy
    weights = np.array([w for _, w in tw])

    # Calculamos la magnitud del vector utilizando NumPy
    magnitude = np.linalg.norm(weights)

    # Verificamos que la magnitud no sea cero para evitar divisiones por cero
    if magnitude != 0:
        # Normalizamos cada peso dividiéndolo por la magnitud utilizando NumPy
        normalized_weights = weights / magnitude
    else:
        # Si la magnitud es cero (todos los pesos son cero), devolvemos el vector original
        normalized_weights = weights

    # Combinamos los términos con los pesos normalizados en una lista de tuplas
    normalized = [(t, w) for t, w in zip([t for t, _ in tw], normalized_weights)]

    return normalized

```

```python

def print_term_weight_vector(twv):
    for term, weight in twv:
        print(f"{term}: {weight}")


def cosine_similarity(tw1, tw2):

    # Obtener los términos comunes entre tw1 y tw2
    common_terms = set([t for t, _ in tw1]) & set([t for t, _ in tw2])

    # Crear listas para almacenar los valores de los términos comunes en tw1 y tw2
    values_tw1 = []
    values_tw2 = []

    # Obtener los valores de los términos comunes y guardarlos en las listas
    for term in common_terms:
        values_tw1.append([w for t, w in tw1 if t == term][0])  # Buscar el valor correspondiente en tw1
        values_tw2.append([w for t, w in tw2 if t == term][0])  # Buscar el valor correspondiente en tw2

    # Calcular el producto escalar entre los vectores normalizados
    dot_product = np.dot(values_tw1, values_tw2)

    return dot_product
```

Exemples amb documents simples (de prova)

Per provar que el nostre codi és correcte, hem fet una prova amb documents simples d'entre una i cinc paraules per veure si les operacions surten correctament:

Primer hem de crear l'índex dels documents anomenat *default_docs:*

```
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 IndexFilesPreprocess.py --index default_docs --pa
th /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs
Indexing 7 files
Reading files ...
/home/marc/.local/lib/python3.8/site-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarn
ing: the default value for the ?wait_for_active_shards parameter will change from '0' to 'index-setting' in vers
ion 8; specify '?wait_for_active_shards=index-setting' to adopt the future default behaviour, or '?wait_for_acti
ve_shards=0' to preserve today's behaviour
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/home/marc/.local/lib/python3.8/site-packages/elasticsearch/connection/base.py:177: ElasticsearchDeprecationWarn
ing: [types removal] Using include_type_name in put mapping requests is deprecated. The parameter will be remove
d in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'default_docs': {'settings': {'index': {'routing': {'allocation': {'include': {'_tier_preferenc
e': 'data_content'}}}, 'number_of_shards': '1', 'provided_name': 'default_docs', 'creation_date': '1696872671228
', 'analysis': {'analyzer': {'default': {'filter': ['lowercase'], 'type': 'custom', 'tokenizer': 'standard'}}},
'number_of_replicas': '1', 'uuid': 'TKrRJXunRuu9ZZE8i0o_NQ', 'version': {'created': '7171299'}}}}}
Indexing ...

real    0m0,997s
user    0m0,482s
sys     0m0,648s
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$
```

Després ho podem provar el fitxer amb els diferents documents que conté *docs* i veure la seva similaritat:

- Amb una sola paraula en comú:

```
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 TFIDFViewer.py --index default_docs --files
/home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d1.txt /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d5.txt
 --print
TFIDF FILE /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d1.txt
tres: 0.17899392988224255
un: 0.9838501781599223
---------------------
TFIDF FILE /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d5.txt
quatre: 0.9471171477160415
sis: 0.3157057159053472
tres: 0.05743700416040321
---------------------
Similarity = 0.01028

real    0m0,514s
user    0m0,556s
sys     0m0,666s
```

- No tenen cap paraula en comú:

```
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 TFIDFViewer.py --index default_docs --files
/home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d4.txt /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d7.txt
 --print
TFIDF FILE /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d4.txt
dos: 0.9349893281889134
sis: 0.3161868913681353
tres: 0.028762272711624595
un: 0.15809344568406766
---------------------
TFIDF FILE /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d7.txt
cinc: 0.8283323365465771
quatre: 0.5602370393246845
---------------------
Similarity = 0.00000

real    0m0,429s
user    0m0,475s
sys     0m0,796s
```

- Tenen totes les paraules en comú: (mateix document)

```
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 TFIDFViewer.py --index default_docs --files
/home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d1.txt /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d1.txt
 --print
TFIDF FILE /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d1.txt
tres: 0.17899392988224255
un: 0.9838501781599223
---------------------
TFIDF FILE /home/marc/Escritorio/REIN/lab2/2REIN-lab/docs/d1.txt
tres: 0.17899392988224255
un: 0.9838501781599223
---------------------
Similarity = 1.00000

real    0m0,647s
user    0m0,522s
sys     0m0,688s
```

Com podem observar el nostre codi funciona correctament, ja que si se li passem uns documents amb les mateixes paraules (el mateix dues vegades) retorna una similaritat de 1, si se li passa uns documents amb cap paraula en comú retorna una similaritat de 0 i si hi han algunes paraules en comú retorna en precisió la similaritat calculada segons la fórmula entre els documents, en aquest cas calcula un 0.01028 de similitud ja que una sola paraula tenen en comú.

Exemples amb documents de novels:

Una vegada provat que el fitxer modificat *TFIDFViewer.py* funciona, ho podem aplicar en el corpus de novel·les i poder veure les similituds entre diferents documents:

Indexar les novel·les utilitzant el token de *whitespace* per separar les paraules, i el filtre de *lowercase* i *asciifolding* per poder fer una filtració que hem trobat addient al tractar de un índex de novel·les en anglès:

```
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 IndexFilesPreprocess.py --index n
ovels_lab2 --path /home/marc/Escritorio/REIN/lab2/2REIN-lab/novels/ --token whitespace --filter
lowercase asciifolding
Indexing 33 files
Reading files ...
/home/marc/.local/lib/python3.8/site-packages/elasticsearch/connection/base.py:177: Elasticsearc
hDeprecationWarning: the default value for the ?wait_for_active_shards parameter will change fro
m '0' to 'index-setting' in version 8; specify '?wait_for_active_shards=index-setting' to adopt
the future default behaviour, or '?wait_for_active_shards=0' to preserve today's behaviour
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
/home/marc/.local/lib/python3.8/site-packages/elasticsearch/connection/base.py:177: Elasticsearc
hDeprecationWarning: [types removal] Using include_type_name in put mapping requests is deprecat
ed. The parameter will be removed in the next major version.
  warnings.warn(message, category=ElasticsearchDeprecationWarning)
Index settings= {'novels_lab2': {'settings': {'index': {'routing': {'allocation': {'include': {'
_tier_preference': 'data_content'}}}, 'number_of_shards': '1', 'provided_name': 'novels_lab2', '
creation_date': '1696965071030', 'analysis': {'analyzer': {'default': {'filter': ['lowercase', '
asciifolding'], 'type': 'custom', 'tokenizer': 'whitespace'}}}, 'number_of_replicas': '1', 'uuid
': 'Ejn5v-vESF-xcmuLqnzrHg', 'version': {'created': '7171299'}}}}}
Indexing ...

real    0m6,765s
user    0m2,332s
sys     0m0,969s
```

Comparació entre diferents novel·les:

- WellsTimeMachine.txt vs LondonCallofTheWild.txt

```
you.": 0.0015847555135046004
young: 0.0006281433007268808
young,: 0.0007923777567523002
younger: 0.0023771332702569
youngster: 0.003028765004147569
your: 0.0
youth: 0.0005181583485564929
ysabel: 0.010036995064922987
ysabel,: 0.010036995064922987
yukon: 0.030110985194768958
yukon,: 0.0050184975324614935
zeal,: 0.0022255548289913853
zero: 0.0034416718105401605
zero): 0.0050184975324614935
zero,: 0.0034416718105401605
--------------------
Similarity = 0.01954

real    0m2,988s
user    0m2,888s
sys     0m0,746s
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 TFIDFViewer.py --index novels_lab
2 --files /home/marc/Escritorio/REIN/lab2/2REIN-lab/novels/WellsTimeMachine.txt /home/marc/Escri
torio/REIN/lab2/2REIN-lab/novels/LondonCallofTheWild.txt --print
```

- PoeWorksVol2.txt vs RiceBurroughsAprinnessofMars.txt

```
zealously: 0.002462588760023385
zitidar: 0.003071478172713466
zitidar,: 0.003071478172713466
zitidars: 0.006142956345426932
zitidars,: 0.012285912690853863
zodanga: 0.06757251979969625
zodanga,: 0.05528660710884238
zodanga.: 0.03685773807256159
zodanga.": 0.003071478172713466
zodangan: 0.05528660710884238
zodangan.: 0.006142956345426932
zodangan?": 0.003071478172713466
zodangans: 0.02764330355442119
zodangans,: 0.009214434518140397
zodangans.: 0.006142956345426932
---------------------
Similarity = 0.05572

real    0m14,356s
user    0m14,188s
sys     0m1,430s
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 TFIDFViewer.py --index novels_lab
2 --files /home/marc/Escritorio/REIN/lab2/2REIN-lab/novels/PoeWorksVol2.txt /home/marc/Escritori
o/REIN/lab2/2REIN-lab/novels/RiceBurroughsAprincessofMars.txt --print
```

- PoeWorksVol2.txt vs PoeWorksVol2.txt

```
youth: 0.0028801940971620076
youth!: 0.005591359793976042
youth,: 0.0024179005916295827
yowling: 0.006973856066338088
zaiat:--"_dicebant: 0.006973856066338088
zaiat_.: 0.006973856066338088
zaire: 0.006973856066338088
zaire,: 0.006973856066338088
zaire.: 0.006973856066338088
zealand,: 0.005591359793976042
zephyrs,: 0.006973856066338088
zig-zag: 0.005591359793976042
zigzag: 0.013947712132676177
zimmerman,: 0.005591359793976042
zohar: 0.006973856066338088
---------------------
Similarity = 1.00000

real    1m1,829s
user    1m1,710s
sys     0m2,564s
marc@marcperezg:~/Escritorio/REIN/lab2/2REIN-lab$ time python3 TFIDFViewer.py --index novels_lab
2 --files /home/marc/Escritorio/REIN/lab2/2REIN-lab/novels/PoeWorksVol2.txt /home/marc/Escritori
o/REIN/lab2/2REIN-lab/novels/PoeWorksVol2.txt --print
```

Com podem comprovar al fer la similitud entre novel·les, podem observar en la primera prova retorna un 0.01954, un valor molt baix entre aquestes, això podría ser perque una novel·la és de tipus fantàstica i l'altre és una història d'aventura. En la segona prova ens retorna un 0.05572 una similitud prou més alta al tractar-se d'un llibre de relats de ficció de Edgar Allan Poe amb una història de ficció de Edgar Rice Burroughs tenint així moltes més paraules en comú .

Finalment per assegurar-nos que el codi funciona correctament, hem comparat la mateixa novel·la de Edgar Allan Poe amb ella mateixa, trigant fins a més de 1 minut cada una de totes les paraules per retornar un valor de 1 en similitud.