Universidad Autónoma de Nuevo León
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos
Alder López Cerda

Resumen

Se realizará un análisis de texto para identificar que palabras sobre salen en un libro.

Como primer paso seleccionamos el libro a analizar, en este caso fue *The Critique of Pure Reason, by Immanuel Kant* (ver fig. 1): https://www.gutenberg.org/ebooks/4280

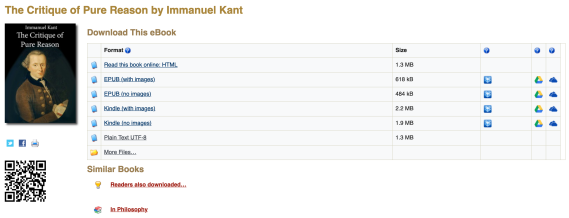


Fig 1

No fue seleccionado por alguna razón en particular, solo con la idea de tener un libro que no fuera técnico, si no que buscara explicar algo en un procesa continua.

Sin afán de meternos a mucho detalla del contenido del libro es posible comentar a modo resumen que el ll libro trata sobre el conocimiento puro y empírico, así como las diferencias, ya que en el resultado veremos como de acuerdo con el contenido las palabras usadas cobraran mucho sentido.

Procesamiento

Se descargo el archivo plain text (ver Fig 2) en formato UTF-8 para un procesamiento mas sencillo.

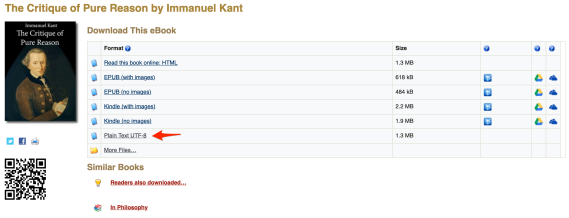
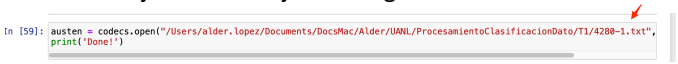


Fig 2

Al colocarse en una ruta local, primero se eliminó el índice, y demás información inicial que no aportarían a nuestro análisis**.**

En nuestro ejercicio se dejó en la sig. ruta:



```
In [59]: austen = codecs.open("/Users/alder.lopez/Documents/DocsMac/Alder/UANL/ProcesamientoClasificacionDato/T1/4280-1.txt",
         print('Done!')
```

Fig 3

Para fines ilustrativos se cambiaron parámetros diversos para mostrar información

```
In [70]: # tokens iniciales
         print(tokens[0:200])
```

```
['Human', 'reason', 'in', 'one', 'sphere', 'of', 'its', 'cognition', 'is', 'called', 'upon', 'to', 'consider', 'que
stions', 'which', 'it', 'cannot', 'decline', 'as', 'they', 'are', 'presented', 'by', 'its', 'own', 'nature', 'but',
'which', 'it', 'cannot', 'answer', 'as', 'they', 'transcend', 'every', 'faculty', 'of', 'the', 'mind', 'It', 'falls
', 'into', 'this', 'difficulty', 'without', 'any', 'fault', 'of', 'its', 'own', 'It', 'begins', 'with', 'principles
', 'which', 'cannot', 'be', 'dispensed', 'with', 'in', 'the', 'field', 'of', 'experience', 'and', 'the', 'truth', '
and', 'sufficiency', 'of', 'which', 'are', 'at', 'the', 'same', 'time', 'insured', 'by', 'experience', 'With', 'the
se', 'principles', 'it', 'rises', 'in', 'obedience', 'to', 'the', 'laws', 'of', 'its', 'own', 'nature', 'to', 'ever
', 'higher', 'and', 'more', 'remote', 'conditions', 'But', 'it', 'quickly', 'discovers', 'that', 'in', 'this', 'way
', 'its', 'labours', 'must', 'remain', 'ever', 'incomplete', 'because', 'new', 'questions', 'never', 'cease', 'to',
'present', 'themselves', 'and', 'thus', 'it', 'finds', 'itself', 'compelled', 'to', 'have', 'recourse', 'to', 'prin
ciples', 'which', 'transcend', 'the', 'region', 'of', 'experience', 'while', 'they', 'are', 'regarded', 'by', 'comm
on', 'sense', 'without', 'distrust', 'It', 'thus', 'falls', 'into', 'confusion', 'and', 'contradictions', 'from', '
which', 'it', 'conjectures', 'the', 'presence', 'of', 'latent', 'errors', 'which', 'however', 'it', 'is', 'unable',
'to', 'discover', 'because', 'the', 'principles', 'it', 'employs', 'transcending', 'the', 'limits', 'of', 'experien
ce', 'cannot', 'be', 'tested', 'by', 'that', 'criterion', 'The', 'arena', 'of', 'these', 'endless', 'contests', 'is
', 'called', '_Metaphysic_', 'Time', 'was', 'when', 'she']
```

```
In [94]: # todo se convierte a minuscula los tokens (palabras) para un procesamiento  uniforme

         words = [word.lower() for word in tokens]

         print(words[:200])
```

```
['human', 'reason', 'in', 'one', 'sphere', 'of', 'its', 'cognition', 'is', 'called', 'upon', 'to', 'consider', 'que
stions', 'which', 'it', 'cannot', 'decline', 'as', 'they', 'are', 'presented', 'by', 'its', 'own', 'nature', 'but',
'which', 'it', 'cannot', 'answer', 'as', 'they', 'transcend', 'every', 'faculty', 'of', 'the', 'mind', 'it', 'falls
', 'into', 'this', 'difficulty', 'without', 'any', 'fault', 'of', 'its', 'own', 'it', 'begins', 'with', 'principles
', 'which', 'cannot', 'be', 'dispensed', 'with', 'in', 'the', 'field', 'of', 'experience', 'and', 'the', 'truth', '
and', 'sufficiency', 'of', 'which', 'are', 'at', 'the', 'same', 'time', 'insured', 'by', 'experience', 'with', 'the
se', 'principles', 'it', 'rises', 'in', 'obedience', 'to', 'the', 'laws', 'of', 'its', 'own', 'nature', 'to', 'ever
```

```
In [99]: # se obtienen las palabras en el texto que no son stop words
         words_ns = [word for word in words if word not in sw]
         print(words_ns[:250])
```

```
['human', 'reason', 'one', 'sphere', 'cognition', 'called', 'upon', 'consider', 'questions', 'cannot', 'decline', '
presented', 'nature', 'cannot', 'answer', 'transcend', 'every', 'faculty', 'mind', 'falls', 'difficulty', 'without
', 'fault', 'begins', 'principles', 'cannot', 'dispensed', 'field', 'experience', 'truth', 'sufficiency', 'time', '
insured', 'experience', 'principles', 'rises', 'obedience', 'laws', 'nature', 'ever', 'higher', 'remote', 'conditio
ns', 'quickly', 'discovers', 'way', 'labours', 'must', 'remain', 'ever', 'incomplete', 'new', 'questions', 'never',
'cease', 'present', 'thus', 'finds', 'compelled', 'recourse', 'principles', 'transcend', 'region', 'experience', 'r
egarded', 'common', 'sense', 'without', 'distrust', 'thus', 'falls', 'confusion', 'contradictions', 'conjectures',
'presence', 'latent', 'errors', 'however', 'unable', 'discover', 'principles', 'employs', 'transcending', 'limits',
'experience', 'cannot', 'tested', 'criterion', 'arena', 'endless', 'contests', 'called', '_metaphysic_', 'time', '_
queen_', 'sciences', 'take', 'deed', 'certainly', 'deserves', 'far', 'regards', 'high', 'importance', 'object', 'ma
tter', 'title', 'honour', 'fashion', 'time', 'heap', 'contempt', 'scorn', 'upon', 'matron', 'mourns', 'forlorn', 'f
orsaken', 'like', 'hecuba', 'modo', 'maxima', 'rerum', 'tot', 'generis', 'natisque', 'potens', 'nunc', 'trahor', 'e
xul', 'inops', 'ovid', 'metamorphoses', 'xiii', 'first', 'government', 'administration', '_dogmatists_', 'absolute
', '_despotism_', 'legislative', 'continued', 'show', 'traces', 'ancient', 'barbaric', 'rule', 'empire', 'gradually
', 'broke', 'intestine', 'wars', 'introduced', 'reign', '_anarchy', '_', '_sceptics_', 'like', 'nomadic', 'tribes',
'hate', 'permanent', 'habitation', 'settled', 'mode', 'living', 'attacked', 'time', 'time', 'organized', 'civil', '
communities', 'number', 'happily', 'small', 'thus', 'could', 'entirely', 'put', 'stop', 'exertions', 'persisted', '
raising', 'new', 'edifices', 'although', 'settled', 'uniform', 'plan', 'recent', 'times', 'hope', 'dawned', 'upon',
'us', 'seeing', 'disputes', 'settled', 'legitimacy', 'claims', 'established', 'kind', '_physiology_', 'human', 'und
erstanding', 'celebrated', 'locke', 'found', 'although', 'affirmed', 'called', 'queen', 'could', 'refer', 'descent
', 'higher', 'source', 'common', 'experience', 'circumstance', 'necessarily', 'brought', 'suspicion', 'claims', '_g
enealogy_', 'incorrect', 'persisted', 'advancement', 'claims', 'sovereignty', 'thus', 'metaphysics', 'necessarily',
'fell', 'back', 'antiquated', 'rotten', 'constitution', '_dogmatism_', 'became', 'obnoxious', 'contempt', 'efforts
', 'made', 'save', 'present', 'methods', 'according', 'general', 'persuasion']
```

Fig 4

Como podemos ver en la sig. figura se modificó algunos parámetros para una mejor ilustración para identificar cuáles son las palabras con mayor frecuencia.

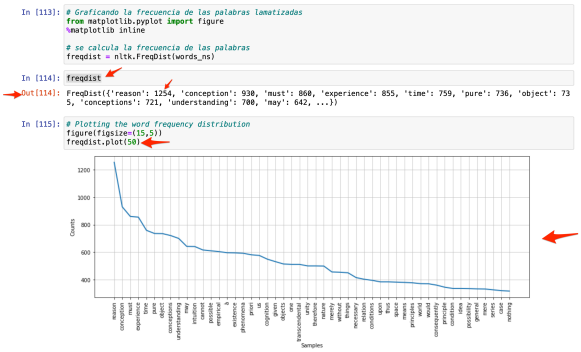


Fig 5

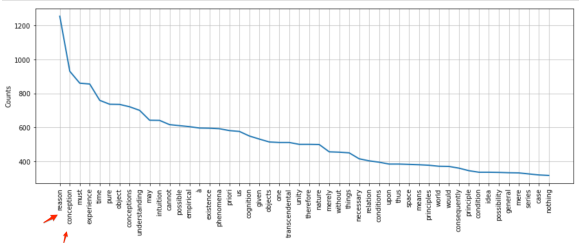


Fig 6

En la *figura 6* indica claramente que los tokens más usados son: *Reason, conception, must, experience, time, pure, object, conceptions, understanding. Lo que nos dice que en el texto hace mucho énfasis a temas como Reason, conception, experience en la redacción.*

Comentarios

Con los tokens que se muestra en el grafico podemos inferir de que se trata el texto, así como que tipo de conversaciones hace mas hincapié el autor.

Adicionalmente

Se trabajo en un ejercicio para obtener el TFIDF de las palabras del libro.

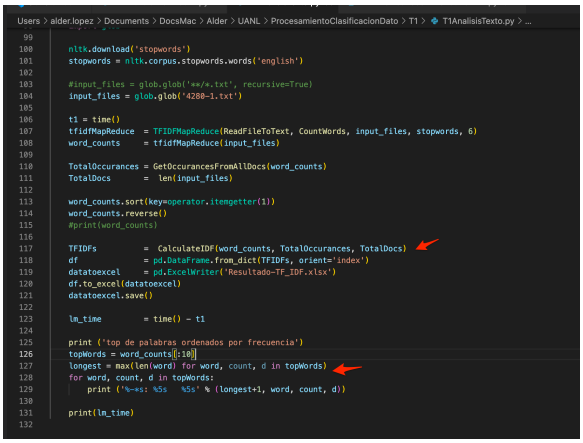
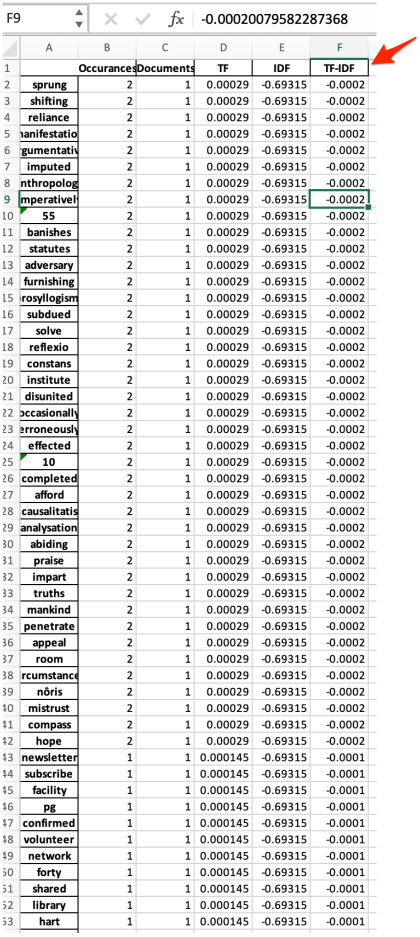


Fig 7



F9 -0.00020079582287368

	A	B	C	D	E	F
1		Occurances	Documents	TF	IDF	TF-IDF
2	sprung	2	1	0.00029	-0.69315	-0.0002
3	shifting	2	1	0.00029	-0.69315	-0.0002
4	reliance	2	1	0.00029	-0.69315	-0.0002
5	anifestatio	2	1	0.00029	-0.69315	-0.0002
6	gumentativ	2	1	0.00029	-0.69315	-0.0002
7	imputed	2	1	0.00029	-0.69315	-0.0002
8	nthropolog	2	1	0.00029	-0.69315	-0.0002
9	mperativel	2	1	0.00029	-0.69315	-0.0002
10	55	2	1	0.00029	-0.69315	-0.0002
11	banishes	2	1	0.00029	-0.69315	-0.0002
12	statutes	2	1	0.00029	-0.69315	-0.0002
13	adversary	2	1	0.00029	-0.69315	-0.0002
14	furnishing	2	1	0.00029	-0.69315	-0.0002
15	rosyllogism	2	1	0.00029	-0.69315	-0.0002
16	subdued	2	1	0.00029	-0.69315	-0.0002
17	solve	2	1	0.00029	-0.69315	-0.0002
18	reflexio	2	1	0.00029	-0.69315	-0.0002
19	constans	2	1	0.00029	-0.69315	-0.0002
20	institute	2	1	0.00029	-0.69315	-0.0002
21	disunited	2	1	0.00029	-0.69315	-0.0002
22	ccasionall	2	1	0.00029	-0.69315	-0.0002
23	rroneously	2	1	0.00029	-0.69315	-0.0002
24	effected	2	1	0.00029	-0.69315	-0.0002
25	10	2	1	0.00029	-0.69315	-0.0002
26	completed	2	1	0.00029	-0.69315	-0.0002
27	afford	2	1	0.00029	-0.69315	-0.0002
28	causalitatis	2	1	0.00029	-0.69315	-0.0002
29	analysation	2	1	0.00029	-0.69315	-0.0002
30	abiding	2	1	0.00029	-0.69315	-0.0002
31	praise	2	1	0.00029	-0.69315	-0.0002
32	impart	2	1	0.00029	-0.69315	-0.0002
33	truths	2	1	0.00029	-0.69315	-0.0002
34	mankind	2	1	0.00029	-0.69315	-0.0002
35	penetrate	2	1	0.00029	-0.69315	-0.0002
36	appeal	2	1	0.00029	-0.69315	-0.0002
37	room	2	1	0.00029	-0.69315	-0.0002
38	rcumstanc	2	1	0.00029	-0.69315	-0.0002
39	nõris	2	1	0.00029	-0.69315	-0.0002
40	mistrust	2	1	0.00029	-0.69315	-0.0002
41	compass	2	1	0.00029	-0.69315	-0.0002
42	hope	2	1	0.00029	-0.69315	-0.0002
43	newsletter	1	1	0.000145	-0.69315	-0.0001
44	subscribe	1	1	0.000145	-0.69315	-0.0001
45	facility	1	1	0.000145	-0.69315	-0.0001
46	pg	1	1	0.000145	-0.69315	-0.0001
47	confirmed	1	1	0.000145	-0.69315	-0.0001
48	volunteer	1	1	0.000145	-0.69315	-0.0001
49	network	1	1	0.000145	-0.69315	-0.0001
50	forty	1	1	0.000145	-0.69315	-0.0001
51	shared	1	1	0.000145	-0.69315	-0.0001
52	library	1	1	0.000145	-0.69315	-0.0001
53	hart	1	1	0.000145	-0.69315	-0.0001

Conclusiones

El análisis de texto tiene una amplia aplicación en revisión de contenido, contexto, sentimientos, y es posible profundizar más en este tema, aunque no está en el alcance de este trabajo.

Github : https://github.com/pacificIT/MCD-Procesamiento-Datos