INSTITUTO POLITECNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO

“Practica 3-Extracción de palabras clave y resumen de textos”

-Tinoco Videgaray Sergio Ernesto

Grupo: 5BV1

Materia: Tecnologías del Lenguaje Natural

01/12/23

- # Introducción.

En esta práctica se van a extraer las palabras clave de 6 documentos correspondientes a diferentes libros con de temas similares, así como también se van a generar los resúmenes de 2 libros de temas distintos de forma automática utilizando diferentes técnicas.

- # Desarrollo

- Extracción de palabras clave.

Generación de cuerpo de documentos.
A través de la pagina https://www.gutenberg.org/ se van a descargar 6 libros, de los cuales 3 pertenecen a temas de tecnología y los otros 3 hablan de psicología (Figura 1).

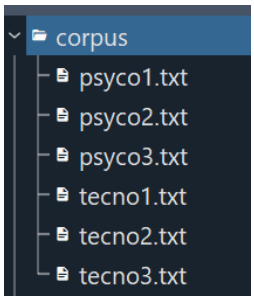

Figura 1. Corpus compuesto por 6 documentos.

- Exploración y limpieza.

Haciendo un análisis de los libros en formato txt se puede observar que hay varios elementos que se pueden omitir como las notas, los anexos y las bibliografías (Figura 2).

```
FOOTNOTES:
[93]The Proceedings and Debates of the 43rd Congress, First Session, 1874 Congressional Rec

[Pg 144]

Bibliography
Adams, Charles K. Sewing machines. In vol. 7 of Johnson's universal cyclopaedia, New York: D

Alexander, Edwin P. Sewing, plaiting, and felting machines. Pp. 341-353 in The Practical Me

——. On the sewing machine: Its history and progress. Journal of the Royal Society of Arts (A

American Historical Society. The life and works of George H. Corliss. (Privately printed for

Biography of Isaac M. Singer. The Atlas, New York, N.Y., March 20, 1853.
```

Figura 2. Notas y Bibliografía colocadas al final del libro.

Para ello se hará uso de expresiones regulares que extraigan el texto que se encuentra antes de los anexos o de la bibliografía (Figura 3).

```
#Extraer texto del libro sin incluir bibliografia o apendices
doc=documento[:re.search("END|APPENDIX|Bibliography|INDEX",documento).start()-3]
```

Figura 3. Expresión regular para omitir bibliografía o apéndices.

- **Extracción de oraciones clave**

Utilizando los algoritmos mostrados a continuación se van a extraer las 5 palabras clave de cada uno de los libros en el corpus.

TF-IDF: Para el algoritmo de TF-IDF que como su nombre lo indica, utiliza los valores TF-IDF de cada documento para establecer un puntaje de los términos clave, se va a realizar una normalización de los documentos que va a consistir en eliminar ciertos signos de puntuación (puntos, comas, paréntesis, corchetes, etc.) y eliminar stopwords.

BERT (Bidirectional Encoder Representations from Transformers): mas que un algoritmo es un modelo ya que utiliza un Transformer el cual va a utilizar diferentes elementos como redes neuronales, autocodificadoras, codificación posicional y vectores de embedding.

Rake (Rapid Automatic Keyword Extraction): utiliza valores como la frecuencia de termino y la co-ocurrencia de términos entre frases, hace la normalización del texto de forma automática y únicamente se debe definir el lenguaje de las stopwords que se vayan a eliminar.

TextRank: utiliza grafos entre los distintos términos para asignar un puntaje a cada termino y frase dentro del documento, así como también la co-ocurrencia entre términos que estén adyacentes, realiza la normalización de forma interna definiendo un “pipeline”.
Primeramente, se van a extraer las 5 mejores palabras clave del primer libro.

- Documento 1.

Para el documento 1 que corresponde con un libro de tecnología destacan palabras como “machines”, “flying machine”, o “helicopter” (Figura 4).

```
Libro Tecnologia1
TF-IDF
['though', 'subject', 'aerial', 'navigation', 'generally']
BERT
[('flying machines', 0.5738), ('flying machine', 0.5634), ('helicopteres', 0.5348), ('helicoptere',
0.4968), ('light fly wheel', 0.4481)]
RAKE
['different recognized authorities vary 50 per cent', '4 ” × 4 ”) would develop', 'tangential force
still came within one pound', 'one mile per hour velocity striking square', 'years ago professor
langley called attention']
TextRank
[('flying machines', 0.08218026534267879), ('machine', 0.0782323709944403), ('machines',
0.0782323709944403), ('large machines', 0.07789478732534005), ('new machines', 0.07714785520944219)]
```

Figura 4. Palabras clave extraídas del primer libro de tecnología.

- Documento 2.

Para el documento 2 que corresponde con un libro de tecnología destacan palabras como “machine”, o “sewing machine” (Figura 5).

```
Libro Tecnologia2
TF-IDF
['preface', 'instrument', 'panel', 'push', 'button']
BERT
[('patent sewing machine', 0.7378), ('modern sewing machine', 0.7368), ('improved patent sewing
machine', 0.7328), ('sewing machine patent', 0.7311), ('important sewing machine', 0.724)]
RAKE
['fine stitching 7 175 binding hats 33 374 stitching vamped shoes 10 210 stitching fine linen 23 640
stitching fine silk 30 550 time', '307 .... centennial ” ” ” .... .... .... .... .... .... 514 ....
.... ....', '665 217 37 .... .... secor ” ” ” .... .... .... .... .... 311 3', '000 .... .... ....
.... .... .... .... .... gold medal ” ” ” .... .... .... 8', 'thompson .... .... .... .... 147 ....
.... .... .... .... union buttonhole machine co']
TextRank
[('Machine sewing', 0.10489878989067486), ('Sewing machines', 0.10489878989067486), ('machine sewing',
0.10489878989067486), ('sewing machine', 0.10489878989067486), ('sewing machines',
0.10489878989067486)]
```

Figura 5. Palabras clave extraídas del segundo libro de tecnología.

- Documento 3.

Para el documento 3 que corresponde con el último libro de tecnología destacan palabras como “steam”, “engine steam”, o “Fig” (Figura 6).

```
Libro Tecnologia3
TF-IDF
['works', 'chapter', 'steam', 'engine', 'the']
BERT
[('steam molecules', 0.6658), ('pressure steam', 0.538), ('pressure steam moves', 0.538), ('steam
turbine', 0.5362), ('boilers', 0.5283)]
RAKE
['ticket – electric train staff system – interlocking – signalling operations – power signalling –
pneumatic signalling – automatic signalling', 'current – sounding instruments – telegraphic relays –
recording telegraphs – high', '1859 .– james bowman lindsay transmitted messages across', 'follows :–
1842 .– professor morse sent aërial messages across', 'expansion engines – compound locomotives –
reversing gears –" linking']
TextRank
[('Fig', 0.048885403596587225), ('engines steam', 0.04858456554403637), ('steam pressure',
0.0477633442725029), ('Fig 117a', 0.04634850175872549), ('other moving parts', 0.04387770064152485)]
```

Figura 6. Palabras clave extraídas del tercer libro de tecnología.

- Documento 4.

Para el documento 4 que corresponde con un libro de psicología destacan palabras como “darwin”, “charles darwin”, o “evolution” (Figura 7).

```
Libro Psicologia1
TF-IDF
['chapter', 'introduction', 'general', 'ignorance', 'subject']
BERT
[('darwinism', 0.5518), ('darwin', 0.5153), ('darwinian', 0.4984), ('charles darwin', 0.4944),
('evolution', 0.4892)]
RAKE
['su tirer avec le temps tous les autres êtres organisés ).', '“ nous ne faisons pas attention que
nous altérons la philosophie ,” “', 'translated “ mais comme nous voulons toujours tout rapporter à un
certain', '“ das biesen des viehes ” affords another example', '“ considérations sur les corps organes
” appeared']
TextRank
[('Mr. Darwin', 0.055936840654596094), ('Mr. Charles Darwin', 0.051966801469822355), ('Dr. Darwin',
0.047110739766761396), ('animal instinct', 0.04684472103456228), ('Dr. Erasmus Darwin',
0.04616038750519434)]
```

Figura 7. Palabras clave extraídas del primer libro de psicología.

- Documento 5.

Para el documento 5 que corresponde con un libro de psicología destacan palabras como “judgment”, “cases”, o “such cases” (Figura 8).

```
Libro Psicologia2
TF-IDF
['part', 'subjective', 'conditions', 'evidence', 'mental']
BERT
[('judgments', 0.4), ('socrates', 0.3873), ('conventional judgment', 0.3842), ('disjunctive judgment',
0.3772), ('priori objections', 0.3743)]
RAKE
['“ nos besoins sont nos forces ”–', 'never testify conscientiously ;– hic niger est', 'physical
correlate ,” says helmholtz ,[ 68', '“ occasiones hominem fragilem non faciunt', 'concealed one ,”
says kraus ,[ 88']
TextRank
[('other cases', 0.049775930657202885), ('Such cases', 0.047241696576840575), ('such cases',
0.047241696576840575), ('Such men', 0.04708190872416293), ('such smaller cases',
0.044582995577016386)]
```

Figura 8. Palabras clave extraídas del segundo libro de psicología.

- Documento 6.

Para el último documento que corresponde con psicología destacan palabras como “mental”, “energy”, o “mental energy” (Figura 9).

```
Libro Psicologia3
TF-IDF
['chapter', 'mental', 'second', 'wind', 'sticking']
BERT
[('mental energy', 0.5794), ('exhaustion', 0.5621), ('potential subconscious reserve energy', 0.5615),
('mind accumulates energy', 0.5577), ('lagging brain', 0.5557)]
RAKE
['great achievers professor walter dill scott', 'represents wasted energy capriciously expended',
'outwardly visible muscular activity expressive', 'call man ," wrote emerson', 'profound problems find
easy solution']
TextRank
[('Mental energy', 0.07834337195686752), ('mental energies', 0.07834337195686752), ('mental energy',
0.07834337195686752), ('latent mental energy', 0.07275042620013877), ('mental power',
0.065250445054254)]
```

Figura 9. Palabras clave extraídas del tercer libro de psicología.

De las 4 técnicas utilizadas, yo considero que las que mejores resultados dieron fueron TextRank y BERT ya que a diferencia de TF-IDF que prácticamente tomaba los primeros términos que aparecen en cada texto o RAKE que toma en cuenta las co-ocurencias ponderadas por las frecuencias, TextRank y BERT se basan en modelos mas complejos como Transformers y grafos con modelos previamente entrenados los cuales tienen mejores resultados.

- **Resumen automático extractivo de texto.**

En esta sección se van a seleccionar 2 libros del corpus que pertenezcan a diferentes temas (tecnología y psicología) y se van a generar los resúmenes del primer capítulo de cada libro utilizando las técnicas que se mencionan a continuación:

Frecuencia de palabras normalizada: se hace la normalización del texto por medio del pipeline de spacy, posteriormente se obtienen las frecuencias de cada termino del texto y se normaliza dividiendo entre la frecuencia mas alta y se utilizan dichos valores para realizar el puntaje de cada frase en el texto.
Gensim: basado en TextRank utiliza grafos para asignar el puntaje de cada termino y frase, se va a definir únicamente el numero de palabras que va a contener el resumen o bien el porcentaje del texto que se va a utilizar.

TextRank: como ya se mencionó anteriormente, TextRank utiliza grafos y la co-ocurrencia para asignar los puntajes entre cada términos y frases dentro del texto del documento. Se va a indicar el numero de frases que se van a utilizar para generar el resumen.

LSA (Latent Semantic Analysis): es una técnica que utiliza elementos algebraicos y estadísticos avanzados para extraer estructuras semánticas que se encuentran dentro del texto de forma implícita. Para su implementación se define ya sea un radio de extracción o bien un numero de frases para generar el resumen de forma similar a como se implementa TextRank.

A continuación, se van a generar los resúmenes correspondientes a cada técnica de las ya mencionadas para el primer capítulo de un libro de Psicología.

- Documento 1

Primeramente, se generan los resúmenes del libro de Psicología utilizando las 10 frases con mejor puntaje asignado por cada modelo (Figura 10-13).

```
Libro Psicologia
Freq Normalizada de Palabras
To quote an example from the last few weeks only, [2] I have observed that Professor Huxley has
celebrated the twenty-first year since the “Origin of Species” was published by a lecture at the Royal
Institution, and am told that he described Mr. Darwin’s candour as something actually “terrible” (I give
Professor Huxley’s own word, as reported by one who heard it); and on opening a small book entitled
“Degeneration,” by Professor Ray Lankester, published a few days before these lines were written, I find
the following passage amid more that is to the same purport

“Suddenly one of those great guesses which occasionally appear in the history of science was given to
the science of biology by the imaginative insight of that greatest of living naturalists—I would say
that greatest of living men—Charles Darwin.I failed to see how important it was that these two theories—
if indeed “natural selection” can be called a theory—should not be confounded together, and that a
“theory of descent with modification” might be true, while a “theory of descent with modification
through natural selection”He did not even, I am assured, mention “natural selection,” but appeared to
believe, with Professor Tyndall, [10a] that “evolution” is “Mr. Darwin’s theory.”  In his later editions
(I believe first in his third, when 6000 copies had been already sold), Mr. Darwin did indeed introduce
a few pages in which he gave what he designated as a “brief but imperfect sketch” of the progress of
opinion on the origin of species prior to the appearance of his own work; but the general impression
which a book conveys to, and leaves upon, the public is conveyed by the first edition—the one which is
alone, with rare exceptions, reviewed; and in the first edition of the “Origin of Species” Mr. Darwin’s
```

Figura 10. Resumen del primer capitulo del libro de psicología utilizando frecuencia normalizada de palabras.

```
Gensim
As a parallel to this, though in respect of the rapid spread of an opinion, and not its decadence, it is
probable that those of our descendants who take an interest in ourselves will note the suddenness with
which the theory of evolution, from having been generally ridiculed during a period of over a hundred
years, came into popularity and almost universal acceptance among educated people.
To quote an example from the last few weeks only, [2] I have observed that Professor Huxley has
celebrated the twenty-first year since the “Origin of Species” was published by a lecture at the Royal
Institution, and am told that he described Mr. Darwin’s candour as something actually “terrible” (I give
Professor Huxley’s own word, as reported by one who heard it); and on opening a small book entitled
“Degeneration,” by Professor Ray Lankester, published a few days before these lines were written, I find
the following passage amid more that is to the same purport
I believe, however, that though we must always gladly and gratefully owe it to Mr. Darwin and Mr.
Wallace that the public mind has been brought to accept evolution, the admiration now generally felt for
the “Origin of Species” will appear as unaccountable to our descendants some fifty or eighty years hence
as the enthusiasm of our grandfathers for the poetry of Dr. Erasmus Darwin does to ourselves; and as one
who has yielded to none in respect of the fascination Mr. Darwin has exercised over him, I would fain
say a few words of explanation which may make the matter clearer to our future historians.
```

Figura 11. Resumen del primer capítulo del libro de psicología utilizando la técnica de gensim.

```
TextRank
I believe, however, that though we must always gladly and gratefully owe it to Mr. Darwin and Mr.
Wallace that the public mind has been brought to accept evolution, the admiration now generally felt for
the “Origin of Species” will appear as unaccountable to our descendants some fifty or eighty years hence
as the enthusiasm of our grandfathers for the poetry of Dr. Erasmus Darwin does to ourselves; and as one
who has yielded to none in respect of the fascination Mr. Darwin has exercised over him, I would fain
say a few words of explanation which may make the matter clearer to our future historians.As far as the
ideas are concerned this is certainly the case, and considering that Professor Hering wrote between
seven and eight years before I did, I think it due to him, and to my readers as well as to myself, to
explain the steps which led me to my conclusions, and, while putting Professor Hering’s lecture before
them, to show cause for thinking that I arrived at an almost identical conclusion, as it would appear,
by an almost identical road, yet, nevertheless, quite independently, I must ask the reader, therefore,
to regard these earlier chapters as in some measure a personal explanation, as well as a contribution to
the history of an important feature in the developments of the last twenty years.We did not know, then,
that according to the earlier writers the variations whose accumulation results in species were not
fortuitous and definite, but were due to a known principle of universal application—namely, “sense of
need”—or apprehend the difference between a theory of evolution which has a backbone, as it were, in the
tolerably constant or slowly varying needs of large numbers of individuals for long periods together,
and one which has no such backbone, but according to which the progress of one generation is always
liable to be cancelled and obliterated by that of the next.[8b]  This should perhaps be a delicately
ironical way of hinting that Mr. Darwin did not read his grandfather’s books closely; but I hardly think
```

Figura 12. Resumen del primer capítulo del libro de psicología utilizando TextRank.

```
LSA
There is no living philosopher who has anything like Mr. Darwin’s popularity with Englishmen
generally; and not only this, but his power of fascination extends all over Europe, and
indeed in every country in which civilisation has obtained footing: not among the illiterate
masses, though these are rapidly following the suit of the educated classes, but among
experts and those who are most capable of judging.Being on my way to New Zealand when the
“Origin of Species” appeared, I did not get it till 1860 or 1861.Buffon we knew by name, but
he sounded too like “buffoon” for any good to come from him.We took it for granted that more
light must be being thrown instead of less; and reading in perfect good faith, we rose from
our perusal with the impression that Mr. Darwin was advocating the descent of all existing
forms of life from a single, or from, at any rate, a very few primordial types; that no one
else had done this hitherto, or that, if they had, they had got the whole subject into a
mess, which mess, whatever it was—for we were never told this—was now being removed once for
all by Mr. Darwin.But perhaps nothing more prepossessed us in Mr. Darwin’s favour than the
air of candour that was omnipresent throughout his work.It never occurred to us that there
might be other and more dangerous opponents who were not brought forward.If they had written
anything worthy of our attention, or indeed if there had been any earlier writers at all,
Mr. Darwin would have been the first to tell us about them, and to award them their due meed
of recognition.Nevertheless, innocent as Mr. Darwin’s opening sentence appeared, it
contained enough to have put us upon our guard.It is observable that the leading men of
```

Figura 13. Resumen del primer capítulo del libro de psicología utilizando LSA.

En el caso del libro de psicología yo considero que prácticamente todos los modelos logran resumir el texto que como tal corresponde con el primer capitulo del libro de psicología ya que como tal el primer capítulo de este libro funciona a manera de introducción de la teoría de Darwin y su desarrollo a lo largo del tiempo. Sin embargo, a la hora de hacer un análisis mas profundo yo considero que TextRank y LSA logran extraer la idea principal de este capitulo que es la ignorancia del termino “evolución” en aquel tiempo.
De igual forma se van a generar 4 resúmenes utilizando las mismas técnicas para un libro de Tecnología (Figura 14-17).

- Documento 2

```
Libro Tecnologia
Freq Normalizada de Palabras
What is steam?—The mechanical energy of steam—The boiler—The circulation of water in a
boiler—The enclosed furnace—The multitubular boiler—Fire-tube boilers—Other types of boilers
—Aids to combustion—Boiler fittings—The safety-valve—The water-gauge—The steam-gauge—The
water supply to a boiler.

The water circulation in a locomotive boiler is—upwards at the fire-box end, where the heat
is most intense; forward along the surface; downwards at the smoke-box end; backwards along
the bottom of the barrel.

OTHER TYPES OF BOILERS.

A steam-boiler has the furnace walls surrounded by water, and its function is to transmit
molecular movement (heat, or energy) through the furnace plates to the water until the point
is reached when steam generates.An injector does not work well if the feed-water be too hot
to condense the steam quickly; and it may be taken as a rule that the warmer the water, the
smaller is the amount of it injected by a given weight of steam.[2] Some injectors have
flap-valves covering the overflow orifice, to prevent air being sucked in and carried to the
boiler.
```

Figura 14. Resumen del primer capítulo del libro de tecnología utilizando la frecuencia normalizada de palabras.

```
Gensim
What is steam?—The mechanical energy of steam—The boiler—The circulation of water in a
boiler—The enclosed furnace—The multitubular boiler—Fire-tube boilers—Other types of boilers
—Aids to combustion—Boiler fittings—The safety-valve—The water-gauge—The steam-gauge—The
water supply to a boiler.
A steam-boiler has the furnace walls surrounded by water, and its function is to transmit
molecular movement (heat, or energy) through the furnace plates to the water until the point
is reached when steam generates.
A further improvement results from increasing the number of tubes (Fig. 5), keeping them all
on the slant, so that the heated water and steam may rise freely.
In a steam-boiler the burning fuel is enclosed either by fire-brick or a "water-jacket,"
forming part of the boiler.
On their way to the funnel the flames and hot gases should act on a very large metal or
other surface in contact with the water of the boiler, in order to give up a due proportion
of their heat.
Fig. 7.—The Babcock and Wilcox water-tube boiler.
The water circulation in a locomotive boiler is—upwards at the fire-box end, where the heat
is most intense; forward along the surface; downwards at the smoke-box end; backwards along
the bottom of the barrel.
In marine boilers of cylindrical shape cross water-tubes and fire-tubes are often employed
```

Figura 15. Resumen del primer capítulo del libro de tecnología utilizando Gensim.

```
TextRank
If steam is let into one end of a cylinder behind an air-tight but freely-moving piston, it
will bombard the walls of the cylinder and the piston; and if the united push of the
molecules on the one side of the latter is greater than the resistance on the other side
opposing its motion, the piston must move.From the furnace f the flames and hot gases rise
round the upper end of the sloping tubes tt into the space a, where they play upon the under
surface of h before plunging downward again among the tubes into the space b.The arrows
indicate further journeys upwards into the space c on the right of a fire-brick division,
and past the down tubes ss into d, whence the hot gases find an escape into the chimney
through the opening e. It will be noticed that the greatest heat is brought to bear on tt
near their junction with uu, the "uptake" tubes; and that every succeeding passage of the
pipes brings the gradually cooling gases nearer to the "downtake" tubes ss.The furnace
gases, after leaving the two furnace flues, are deflected downwards into the channel a, by
which they pass underneath the boiler to a point[Pg 26] almost under the furnace, where they
divide right and left and travel through cross passages into the side channels bb, to be led
along the boiler's flanks to the chimney exit c. By this arrangement the effective heating
surface is greatly increased; and the passages being large, natural draught generally
suffices to maintain proper combustion.There is—through the air-doors at the bottom of the
furnace, the furnace itself, and the fire-tubes; and on the way oxygen combines with the
carbon of the fuel, to form carbon dioxide.As soon as the steam pressure on the face of the
```

Figura 16. Resumen del primer capítulo del libro de tecnología utilizando TextRank.

```
LSA
If confined in an air-tight vessel, the molecules have their flights curtailed, and beat
more and more violently against their prison walls, so that every square inch of the[Pg 14]
vessel is subjected to a rising pressure.Having thus partly got their liberty, the molecules
become less active, and do not rush about so vigorously.But our boiler must not be so
wasteful of the heat as is that useful household fixture.Other well-known water-tube boilers
are the Yarrow, Belleville, Stirling, and Thorneycroft, all used for driving marine
engines.The furnace in every case is surrounded by water, and fed through a door at one
side.In marine boilers of cylindrical shape cross water-tubes and fire-tubes are often
employed to increase the heating surface.We may now turn our attention more particularly to
the chemical process called combustion, upon[Pg 27] which a boiler depends for its
heat.Professor Thurston, an eminent authority on the steam-engine, has estimated that a
plain cylindrical[Pg 32] boiler carrying 100 lbs.In either case a rupture[Pg 34] would not
be attended by an explosion, as water is very inelastic.As their mechanism is much the same
as that of ordinary force pumps, which will be described in a later chapter, we may pass at
once to the injector, now almost universally used on locomotive, and sometimes on stationary
boilers.
```

Figura 17. Resumen del primer capítulo del libro de tecnología utilizando LSA.

En el caso del libro 2 que corresponde con un libro de tecnología, considero que los 4 algoritmos logran extraer frases relevantes del primer capitulo ya que capturan la esencia de este capítulo que es explicar que es y como funcionan los mecanismos de vapor. En el caso de Gensim y la frecuencia normalizada de palabras se toman frases del comienzo del capitulo que es donde se plantea la pregunta principal de este capitulo y donde se explican los conceptos clave como los mecanismos de vapor, las calderas y las calderas de locomotora, así como otros tipos de calderas. Y en el caso de LSA y TextRank se seleccionan frases similares, aunque no incluyen la pregunta principal del capitulo y de alguna forma se complementan con los primeros 2 algoritmos. Por lo que yo considero que de alguna forma estos algoritmos tienen enfoques distintos.

Análisis y conclusiones

En esta práctica se realizaron varias técnicas de extracción de palabras clave y generación de resúmenes a partir de las frases con mejor puntaje. Aunque algunos de estos algoritmos ya no se utilicen tanto o que hayan quedado obsoletos con la llegada de nuevos modelos como los Transformers, pueden resultar de utilidad cuando no se cuentan con los recursos suficientes para utilizar estos modelos tan sofisticados.

Muchas de estas técnicas son muy similares, pues la mayoría asigna un puntaje ya sea a las palabras clave o bien a las frases del texto y partir de ahí se puede resumir el texto.

De igual forma es importante realizar un análisis de estos algoritmos, ya que puede ser necesario normalizar el texto como el caso del TF-IDF o la frecuencia de palabras normalizada, aunque otros modelos como BERT o TextRank ya realicen dicho proceso de normalización de forma automática.

Yo considero que una buena práctica seria implementar varias de estas técnicas a la vez para tener más de un modelo de referencia a la hora de realizar una técnica de procesamiento de lenguaje natural y así comparar resultados entre distintos modelos o algoritmos.