Reporte de Análisis Exploratorio de Datos (NLP) - Corpus CNN

Integrantes: Joan Antonio Lazaro Silva, Caro Pérez Horacio, Joaquín Alfredo Castro Córdova
Fecha: 20 de enero de 2026
Corpus: cnn_articles.txt

1. Caracterización General del Corpus

Métrica	**Valor Calculado**
Número total de documentos	2,065,892
Total, de palabras	60,490,380
Longitud del vocabulario	809,367 palabras únicas
Densidad Léxica	1.34%
Promedio de longitud	29.28 palabras por documento
Idioma	Inglés
Tono	Formal / Periodístico

Interpretación de las Métricas

- **Densidad Léxica (1.34%):** Este valor indica un vocabulario altamente repetitivo. En un corpus de más de 60 millones de palabras, una densidad tan baja es típica de textos periodísticos estandarizados donde predominan las palabras funcionales.
- **Extensión de los Documentos:** Con un promedio de **29.28 palabras**, se concluye que el corpus está compuesto por fragmentos breves, titulares o "leads" de noticias, lo que lo hace ideal para tareas de generación de encabezados.

2. Análisis Cualitativo del Corpus (CNN)

- **Número de documentos:** 2,065,892 documentos. Al no contar con categorías explícitas, se realizó un muestreo estadístico que identificó:
    - **Política:** 251,786 documentos.
    - **Deportes:** 140,036 documentos.
    - **Salud/Ciencia:** 99,676 documentos.
    - **Otros:** 1,574,394 documentos.
- **Tema(s) principal(es):** El corpus es de **Noticias Globales**. Los temas predominantes son la política internacional y nacional de EE. UU., seguidos por eventos deportivos y actualidad en salud/ciencia.
- **Tono predominante: Formal y Periodístico.** El uso de vocabulario estandarizado y la ausencia de jerga o lenguaje coloquial en el Top 10 confirman un tono profesional propio de una agencia de noticias.
- **Idioma: inglés.** (Confirmado por el vocabulario único y el análisis de palabras más frecuentes).
- **¿Qué tan "sucio" está el corpus?:** Se considera **Moderadamente Sucio**.
    - Contiene marcas de formato técnicas como @delimiter.
    - Presencia de caracteres extraños y errores de formato como guiones dobles (--) al inicio de miles de líneas.
    - No presenta exceso de HTML, pero sí requiere limpieza de conectores (stopwords) para ser útil.
- **Hallazgos curiosos:**

1. **La etiqueta @delimiter:** Aparece casi 100,000 veces, funcionando como un "muro" entre noticias.

2. **Brevedad Extrema:** El promedio de 29 palabras indica que el corpus no guarda artículos completos, sino fragmentos informativos o "noticias de último minuto".

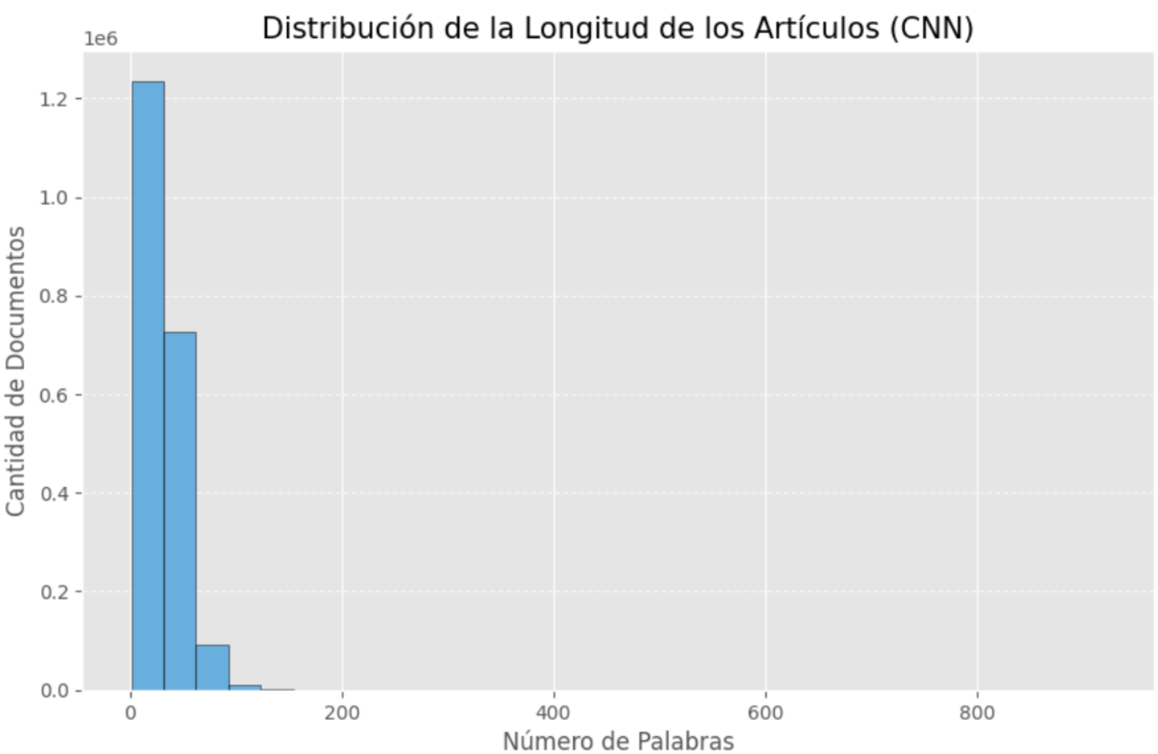
3. **Densidad Léxica Crítica:** Un 1.34% es una de las densidades más bajas posibles, lo que demuestra que el lenguaje periodístico es extremadamente eficiente y repetitivo.

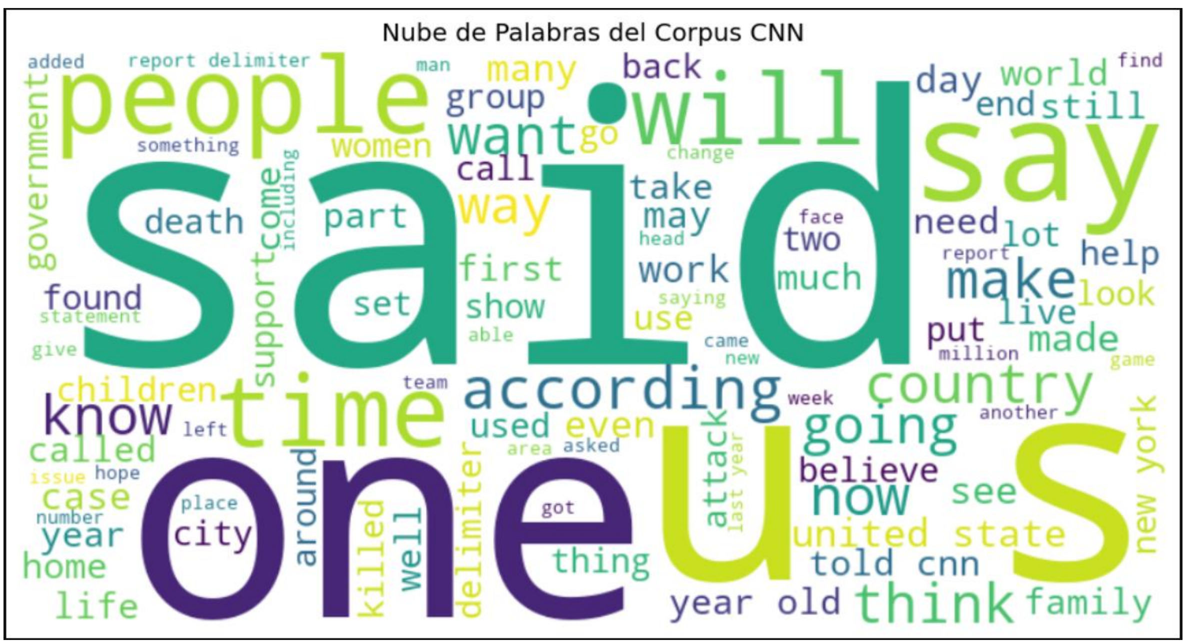
Top 10 palabras más usadas:

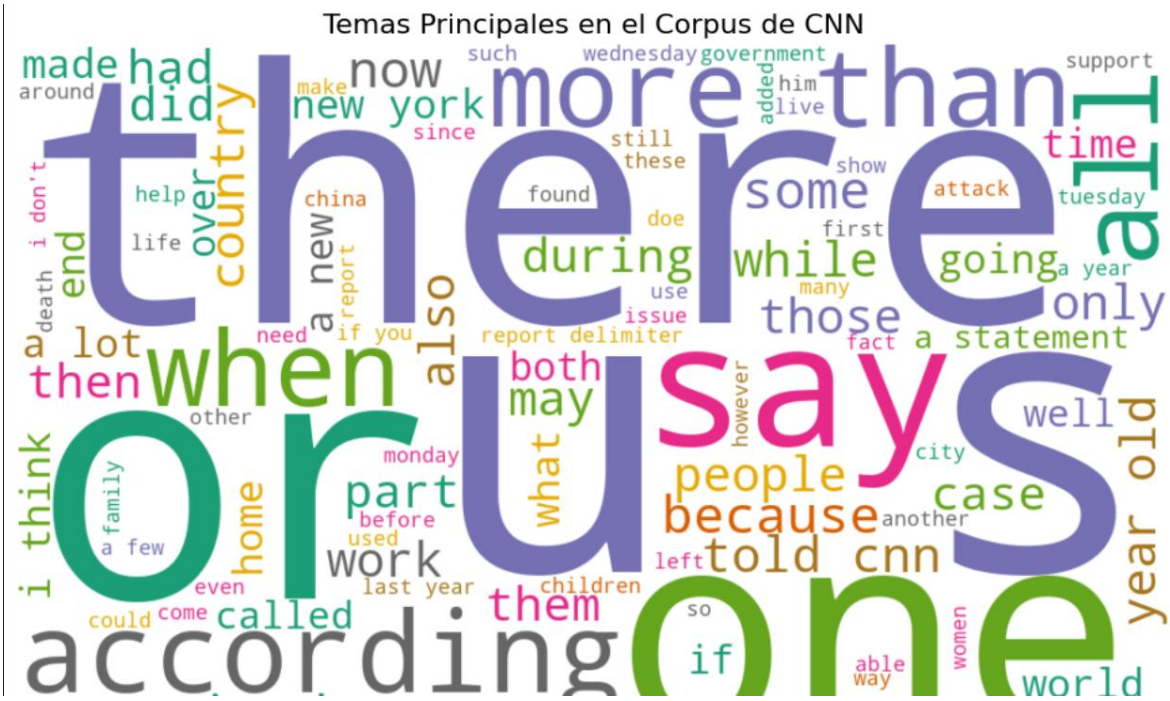
- ('the', 3548361)
- ('to', 1676356)
- ('of', 1514417)
- (‘and’ 1453054)
- ('a', 1437358)
- ('in', 1280296)
- ('that', 652277)
- ('for', 553435)
- ('is', 522919)
- ('on', 475731)

- **Tareas de NLP recomendadas:**
    - **Clasificación de Texto:** Para etiquetar automáticamente el 76% de noticias "sin tema" en categorías específicas.
    - **Generación de Titulares:** Debido a la brevedad de los documentos, es ideal para entrenar modelos que resuman información en una sola frase.
    - **Análisis de Entidades (NER):** Para extraer nombres de personas y lugares geográficos en un flujo masivo de noticias.

Gráficos:


Distribución de la Longitud de los Artículos (CNN)
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Cantidad de Documentos
0
200
400
600
800
Número de Palabras


Nube de Palabras del Corpus CNN
said
one
u
s
people
will
say
time
according
country
know
going
think
make
now
want
way
government
death
support
come
part
first
show
work
much
find
world
day
end
still
many
back
group
women
call
take
may
two
need
lot
help
look
found
set
use
live
put
made
children
called
case
year
city
home
life
around
killed
well
delimiter
thing
attack
used
even
believe
united state
told cnn
year old
family
see
new york
added
report delimiter
man
something
change
including
face
head
report
saying
statement
give
able
came
new
million
game
team
week
another
left
area
asked
issue
hope
place
got
number
last year


Temas Principales en el Corpus de CNN
there
more
than
all
says
or
us
one
according
when
also
both
may
what
part
because
people
told
cnn
work
called
them
if
world
year old
well
case
only
those
during
while
going
some
time
made
had
did
now
new york
country
over
end
home
a new
a lot
then
i think
a few
family
other
monday
before
used
left
another
city
however
fact
a statement
a year
tuesday
attack
show
found
china
help
life
i don't
death
report
need
if you
report delimiter
issue
use
many
doe
first
still
these
since
make
around
such
wednesday
government
added
him
live
support
last year
children
even
could
come
so
women
able
way