Desarrollo de métodos de corrección automática de errores de transcripción de imágenes a texto en expedientes digitales

Comisión Nacional de Búsqueda / CIMAT Monterrey

Alumno: Jorge Luis Vargas Barrera

Encargados:

Dra. Mariana Esther Martinez-Sanchez

Dr. Victor Muñiz

01 Problema

02 Objetivo

03 Método 1

04 Método 2

05 Método 3

06 Siguientes pasos

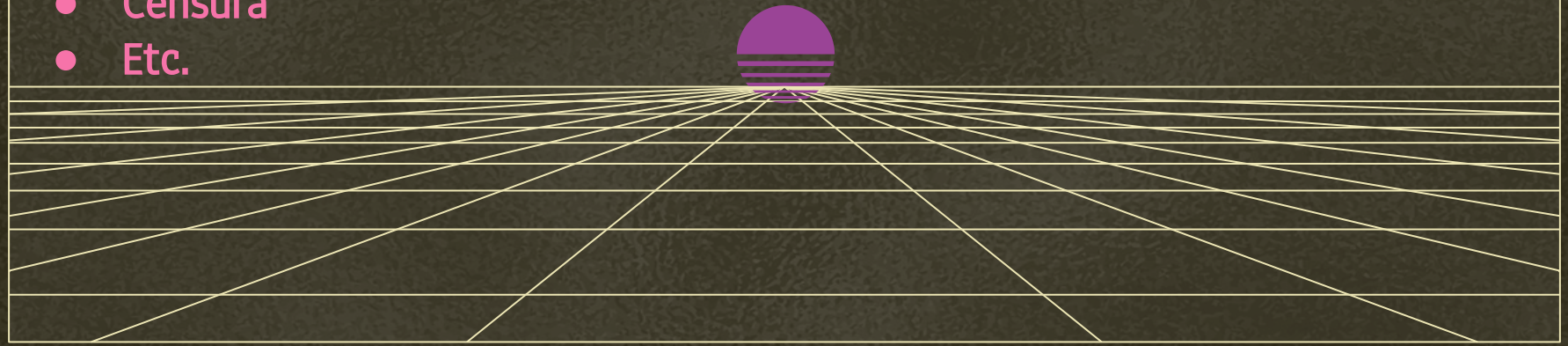
Problema

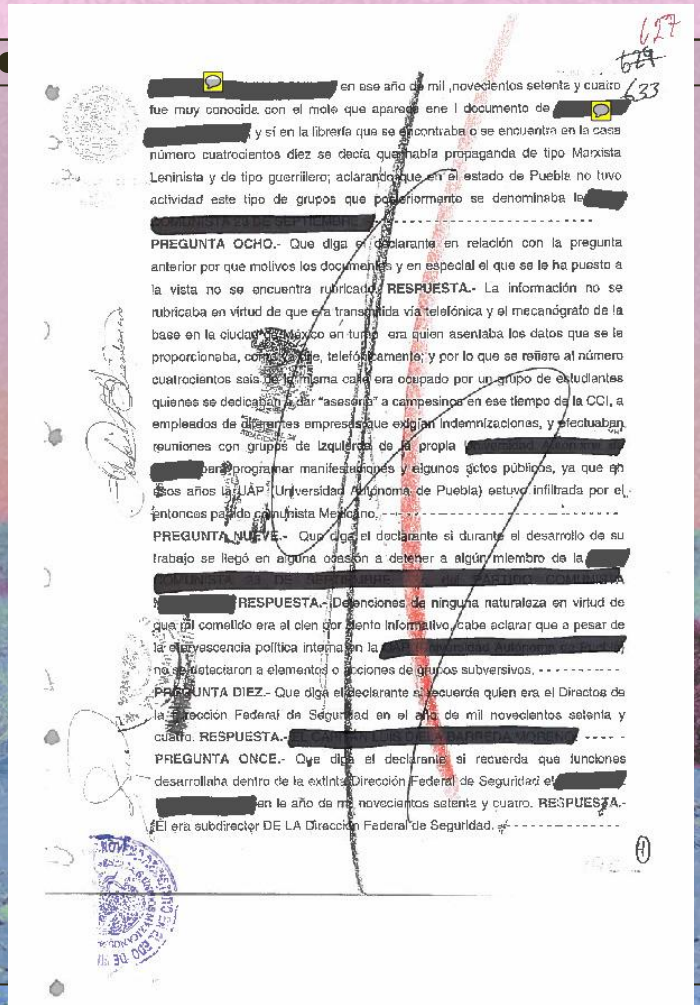
Descripción de la problemática

Se cuenta con métodos basados en OCR para extracción de texto en documentos digitalizados.

Características de documentos:

- Antiguos
- Con diferentes tipografías
- Censura
- Etc.





en ese año de mil ,novecientos setenta y cuatro fue muy conocida con el mote que aparecía ene l documento de [redacted] y si en la librería qua se encontraba o se encuentra en la casa número cuatrocientos diez se decía quehabía propaganda de tipo Marxista Leninista y de tipo guerrillero; aclarandoque en el estado de Puebla no tuvo actividad este tipo de grupos que posteriormente se denominaba le[redacted]

PREGUNTA OCHO.- Que diga el declarante en relación con la pregunta anterior por que motivos los documentos y en especial el que se le ha puesto a la vista no se encuentra rubricado. **RESPUESTA.**- La información no se rubricaba en virtud de que era transmitida vía telefónica y el mecanógrafo de la base en la ciudad de México en turno era quien asentaba los datos que se le proporcionaba, como sabe, telefónicamente; y por lo que se refiere al número cuatrocientos seis de la misma calle era ocupado por un grupo de estudiantes quienes se dedicaban a dar "asesoría" a campesinos en ese tiempo de la COI, a empleados de diferentes empresas que exigían indemnizaciones, y efectuaban reuniones con grupos de izquierda de la propia [redacted] para programar manifestaciones y algunos actos públicos, ya que en esos años la CUAP (Universidad Autónoma de Puebla) estuvo infiltrada por el entonces partido comunista Mexicano.

PREGUNTA NUEVE.- Que diga el declarante si durante el desarrollo de su trabajo se llegó en alguna ocasión a detener a algún miembro de la [redacted] **RESPUESTA.**- Detenciones de ninguna naturaleza en virtud de que tal cometido era al cien por ciento informativo, cabe aclarar que a pesar de la efervescencia política internacional [redacted] no se detectaron a elementos o acciones de grupos subversivos.

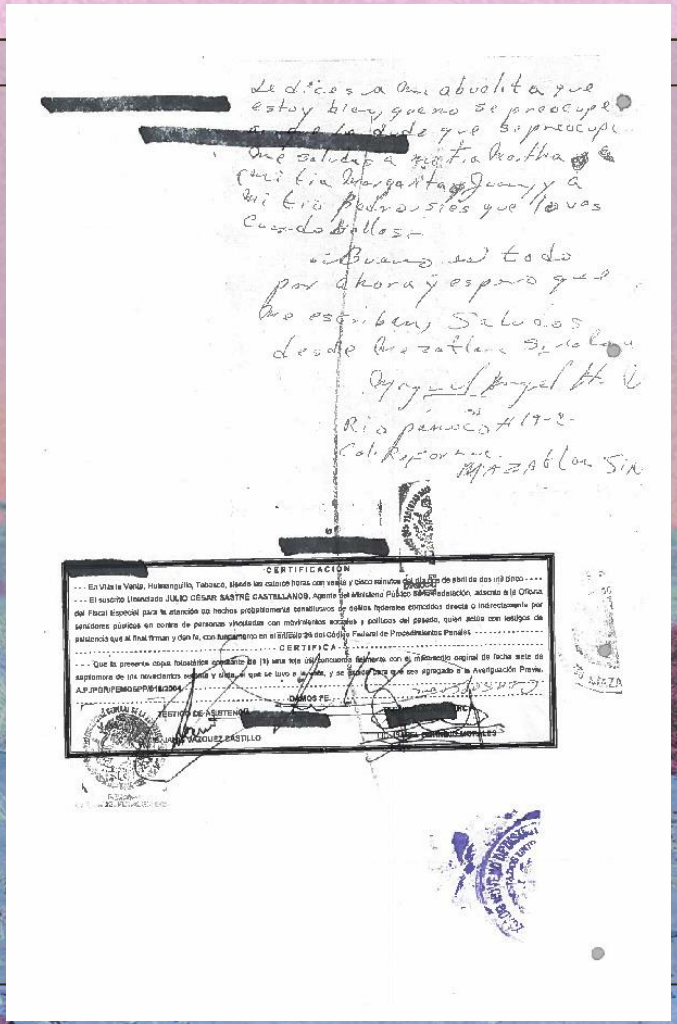
PREGUNTA DIEZ.- Que diga el declarante si recuerda quien era el Directos de la Dirección Federal de Seguridad en el año de mil novecientos setenta y cuatro. **RESPUESTA.**- [redacted]

PREGUNTA ONCE.- Que diga el declarante si recuerda que funciones desarrollaba dentro de la extinta Dirección Federal de Seguridad el [redacted] en le año de mil novecientos setenta y cuatro. **RESPUESTA.**- El era subdirector DE LA Dirección Federal de Seguridad.

---

page-1248.jpg.txt

```
pagina1
¿qe
ns
"E A iaa >> ese año af mil novecientos setenta y cuatro (33
¿52 7. fue muy conocida con el mote que aparegp ene | documento de
00 GA y +0 la oros quo so dicontraba o se encuentra en la casa
Y "2220 húmero cuatrocientos diez se decía queghabía propaganda de tipo Mandsta
De Leninista y de tipo guerillero; aclarandafque ef Al estado de Puebla no tuvo
actividad este tipo de grupos que pasiefíormente se denominaba dy
PREGUNTA OCHO.- Que diga plidiblarante, en relación con la pregunta
anterior por que motivos los docMmehfés y en especial el que se le ha puesto a
la vista no se encuentra SA RESPUESTA La información no se
/%  "ubricaba en virtud de que a/a tran pi ida vigelefónica y el mecanágrao dela
) : x base en la ciuda :0 en tu 1 era gon asentaba los datos que se le
$ y Poporcionaba, A soto am por lo que se refiere al número
o cuatrocientos sois-88 1 risma cali era ocupado por yy río de Syudiantes
NO quienes se dedicalan fiar "asesgrfp" a campesingeén ese tiempo dla COI, a
» ES empleados de iares eromfpo oxojaríindemnizaciones, y Pfectuaban
3| reuniones con grupps de izquidrda de J4 propia
3 GU «8proratrar manites alfiapés Y algunos getos públicbs, ya que eh
|, d30s años I3¿LAP [Unfversidad AMfónoma:de Puebla) estuva infiltrada por el.
: cdo O O
PREGUNTA,NUEVE.- Quydlggel deci te si duranté el desarrollo de su
trabajo se llegó en alyúña ogasión a detgher a algún/miembro de la
vB) pespuesta. aL 'a naturaleza en virtud de
* ques cometio era al cien gor gfonto inforifíaivo cabe aclarar que a pesar de
¿2% WE ofezyescencia política integnafín la
X 5. ¿n8sctectaron a elementos o fbclones de tubos subversivos. -========-
e o INTA DIEZ.- Que diga elfgleciarante Sfecuerda quien era el Directos de
nas a rección Federal de Sdgurblad en el ah de mil novecientos setenta y
y .  cuafto, RESPUESTA. Los a
Y PREGUNTA ONCE. Oyo didk el declkrantél si recuerda que funciones
Y — desarrollaba dentro de la extintafDirección [Federdl de Seguricad e,
— Y A le año de mf novecieftos sotenta y cuatro. RESPUESHA.-
¡4 Él era subdirector DE LA Direc: ign Federal'de Seguridad. $ -==-=-=======-
- -_-_-_—
ee
ELE
RS
mi 300357
Ó
```



page-1099.jpg.txt

```
pagina1
144)
e, que el referido Servidor Público se encontraba en ejercicio de sus funciones po
E en septiembre de mil novecientos setenta y siete época de la detención y 5Sa
0 cesación 6 O AR
ES OTE >.<. 0: "ses py –
AROCURADURIA GENER apareciendo en el reverso. de la última foja la leyenda con lo siguiente: El
pepuytica DIRECTOR GENERAL DE RECURSOS HUMANOS DE LA SECRETARÍA
DE GOBERNACIÓN, CON FUNDAMENTO EN EL ARTÍCULO 9"
FRACCIÓN XIII DEL REGLAMENTO. INTERIOR DE ESTA SECRETARIA,
CERTIFICA: QUE EL PRESENTE DOCUMENTO ES COPIA FIEL Y
CONCUERDA EN TODAS CADA UNA DE SUS PARTES CON EL
ORIGINAL CON EL QUE FUE COTEJADO EL CUAL CONSTA DE
DIECISIETE FOJAS UTIVES Y DEL QUE QUEDA CONSTANCIA EN LOS
ARCHIVOS DE ESTA PIRECCIÓN, SE EXTIENDE A LOS VEINTINUEVE
DÍAS DEL MES DE sí IEMBRE DE DOS MILCUATRO. ========o==..
== - 2, El quince de agosto de dos mil cinco se Tecibió y agregó a la presente
indagatoria el de fecha ocho de
agosto de dos mil ginco, signado por el Subdirebtor de Relaciones Laborales
de la Secretaría Gobernación, mediante ef cual remitió en cinco copias
féfícadas los rjombramien) A MA
informando además su Historial Laboral del cual se
d E ide que Ingresó /Á laborar a la Dirección Federal de Seguridad el
PEfisgo de agosto de pfil novecientos setefta y seis renunciando el dieciséis
dECAREE mil novegiéntos setenta y ocho, reingresando el dieciséis de mayo
de mil novecientos setenta y nueve y renunciando nuevamente por término
Ne¿Ba gfecto de nombramiento el primero de agosto de mil novecientos ochenta
Hey ads, de lo que se desprende que este servidor público se encontraba en
3 ye siglo de sus funciones en septiembre de mil novecientos setenta y siete
e enla que se inició la ejegución del delito de MMMM.
$ 7 en agravio de MAA
e y Documentos
Es A te de cinco fojas útiles, apareciendo en el reverso de la última foja la
ns "léyenda con lo siguiente: El DIRECTOR GENERAL DE RECURSOS
HUMANOS DE LA SECRETARÍA DE" GOBERNACIÓN, CON
. 1436
_Á > -_ — a QA
```

PROCURADURÍA GENERAL DE JUSTICIA

1068 1672

185

Recursos Humanos de la Coordinación Administrativa
13 de Octubre de 2004
Culiacán, Sinaloa. [redacted]
Asunto: Se informa sobre personal.

C. Alfonso Rodriguez Cardenas
Secretario Particular del C. Procurador General de Justicia del Estado
Presente.

En atención a su control de acuerdos [redacted] 26 de Mayo de 2004, por este conducto anexo a Ud, información enviada por la Dirección de Recursos Humanos de Gobierno del Estado mediante oficios No. [redacted] de fecha 13 de [illegible] y 21 de Septiembre respectivamente del presente año, la cual consta de trayectorias laborales del Director de Policía Judicial así como del personal activo correspondiente al año de 1977.

Sin más por el momento, aprovecho la oportunidad para enviarle un cordial saludo.

Atentamente

Leopoldo Ibarra Medina
Coordinador Administrativo

C.c.p. Oscar Fidel González Mendivil.- Procurador General de Justicia del Estado.- Presente.
C.c.p. Archivo.
LIM/AEP/Asj
PGJE/AA/OFICIARIOS

Blvd. Enrique Sánchez Alonso 1833 Nte., Desarrollo Plan 3 Ríos,
C.P. 80030, Culiacán, Sinaloa, México.
Tels. (667) 716-11-20 y 716-11-35 Fax: (667) 713-32-00
E-mail: pgjesin@ges.gob.mx

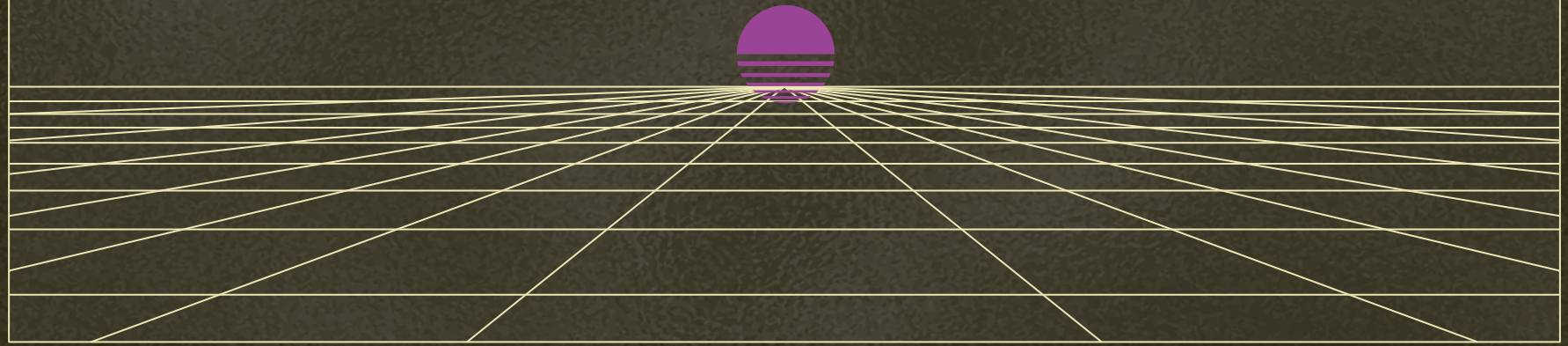
Sinaloa
Con rumbo firme

page-367.jpg.txt

pagina1

"Cómo armó la CNDH el rompecabezas", en Proceso No. 1309, 2 de diciembre de 2001, pp. 34-37.
MORENO BORBOLLA, José Luis. Entrevista a Jaime García Chávez (miembro de la red urbana del Grupo Popular Guerrillero "Arturo Gámiz" comandado por Oscar González Eguiarte), en Para Romper el Silencio Expediente Abierto, Centro de Investigaciones Históricas de los Movimientos Armados, A.C., noviembre 1994-enero 1995, pp. 29-37.
NAVARRO ALTAMIRANO, Juan. "Guadalajara: de Maciel a Parrés. La nueva doctrina universitaria", en Por que? Revista independiente No. 199, abril 20 de 1972. pp. 20-21.
"Guadalajara ¡A seguir la lucha!", en Por qué? Revista independiente No. 218, agosto 31 de 1972. pp.12-14.
"Guadalajara. La tarea: fijar a traidores y oportunistas", en Por qué? Revista independiente No. 183, diciembre 30 de 1971. pp. 25-27.
ORTEGA JUÁREZ, Joel. "El 68: punto final a las fantasías, en Campus Milenio No. 377, jueves 22 de julio de 2010, pp. 8-9.
PONCE DE LEÓN, Alberto. "El hombre de los 200 asesinatos", en Proceso No. 1361, 1" de diciembre de 2002, pp. 12.
Por esto! No. 68, 14 de octubre de 1982, "Dolorosos alto en el camino ¡Al rescate de Por esto!
Punto Crítico No. 4, abril de 1972, "Chihuahua: la verdad sobre los asaltos" pp. 19-21.
Punto Crítico No. 10, Octubre de 1972, "Tribunal Popular. La Sentencia" pp. 32-33
Punto Crítico No. 69, 31 de Enero de 1977, "Denuncian represión en Chihuahua", pp. 38-40.
Punto Crítico, Revista de Información y Análisis Político, No. 50, primera quincena de abril de 1976, "Carta Abierta al Pueblo de México de los padres de los Gámiz", p. 4.
Punto Crítico No. 28, mayo de 1974, "Comunicado de prensa, presos políticos de la crujía "M",p.3,
Punto Crítico, No. 19, agosto de 1973, p. 23.
Punto Crítico, No. 52, primera quincena de mayo de 1976, "Guerrilleros ajustan cuentas con su pasado", pp. 15-18
Punto Crítico No. 32, Julio de 1975, "En C.U. la Ley del Revolver", p. 5.
Punto Crítico, No. 91, Octubre de 1978, "La amnistía una caricatura", pp.6 -7.
Punto Crítico No. 67, 15 de Noviembre de 1976, "Naucalpan, ¿Estado de México o de sitio? pp. 14-15.
Punto Crítico, No. 20/21, septiembre-octubre de 1973, "Los Presos Políticos ante los últimos secuestros. No acepto dicha liberación. Al pueblo de México, pp. 1-2.
Punto Crítico, No. 16, abril de 1973, "La Federación Estudiantil Oaxaqueña denuncia el cacicazgo de Jamiltepec, p. 1.
RAMÍREZ LADEWIG, Carlos. "Indignidad de los diputados", en Proceso, No. 14, 5 de febrero de 1977, p. 8.

- La calidad de los textos extraídos es deficiente en ocasiones.
- Se cuenta con un método para evaluar la calidad de la extracción.



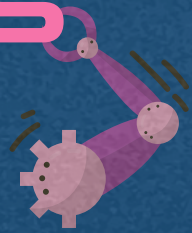
Objetivo

Se pretende desarrollar una serie de métodos para la corrección automática de errores ortográficos/gramaticales.

Basados en:

- ➔ Transformaciones morfológicas de palabras basadas en vocabularios.
- ➔ Métodos probabilísticos
- ➔ Arquitecturas de aprendizaje profundo (redes recurrentes, convolucionales y transformers)

Conjunto de datos



Conjunto de datos

Se tomaron textos con buena calidad de visualización. Los cuales consisten de tesis que

Método 1

método probabilístico

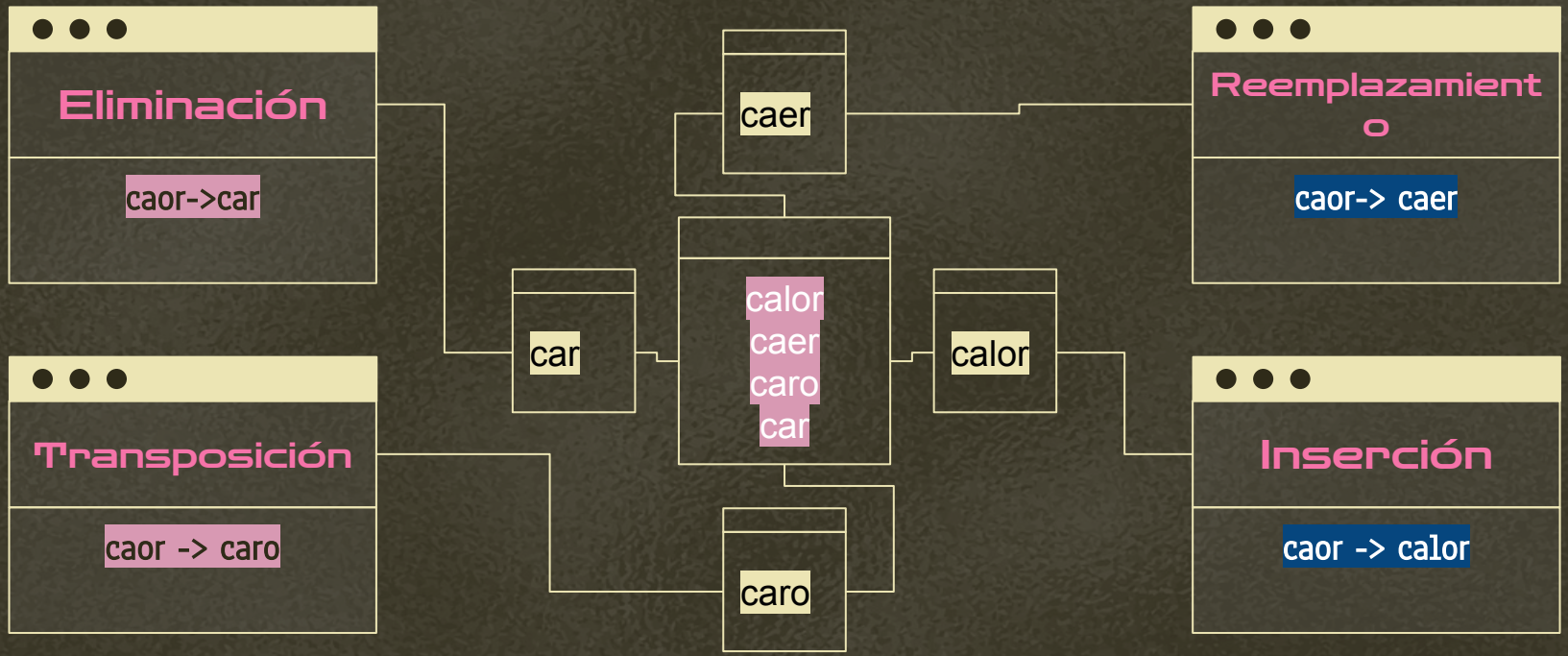
Se utiliza el principio de máxima verosimilitud. Dada una palabra incorrecta s, se busca el mejor candidato w_b que corrija de una lista de posibles candidatos $C(s)$ con la mayor probabilidad (sin normalizar)

$$w_b = arg \max_{w_i \in C(s)} P(s|w_i)P(w_i)$$

Componentes

1. Diccionario: Se detectan los errores de escritura y propone candidatos
2. Modelo de error: $P(s|w_i)$ la probabilidad de que s sea escrita en un texto cuando el autor se refería a w_i
3. Modelo de lenguaje: $P(w_i)$ la probabilidad de que w_i aparezca como una palabra en el texto

Generación de candidatos

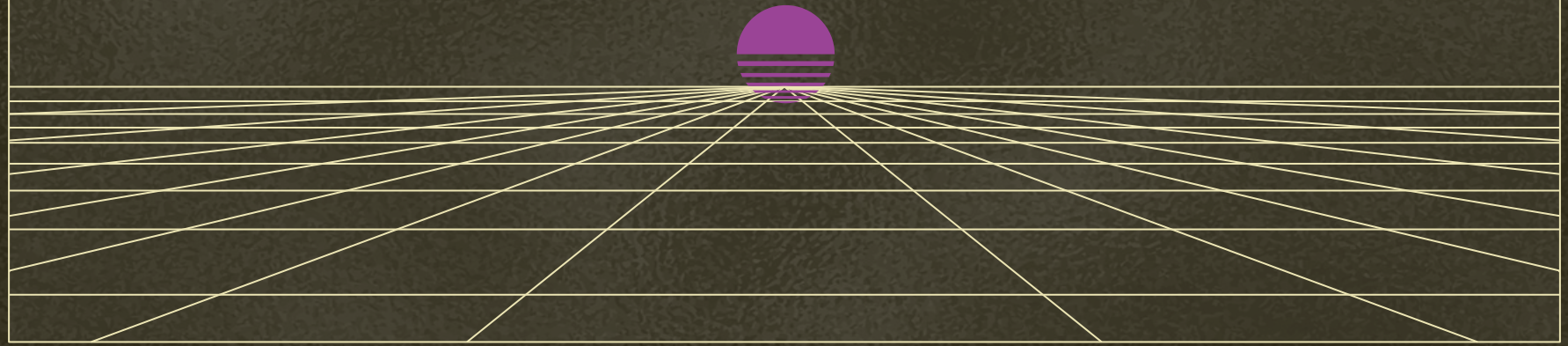


ORACION ORIGINAL
gobernadores, todos del mismo paktid0, los pkesidente8

ORACION CORREGIDA
gobernadores todos del mismo partido, los presidente

ORACION ORIGINAL
pke8idente utilizaba. Bn otro nivel de la pikámide estaban lus

ORACION CORREGIDA
presidente utilizaban en otro nivel de la pirámide estaban los



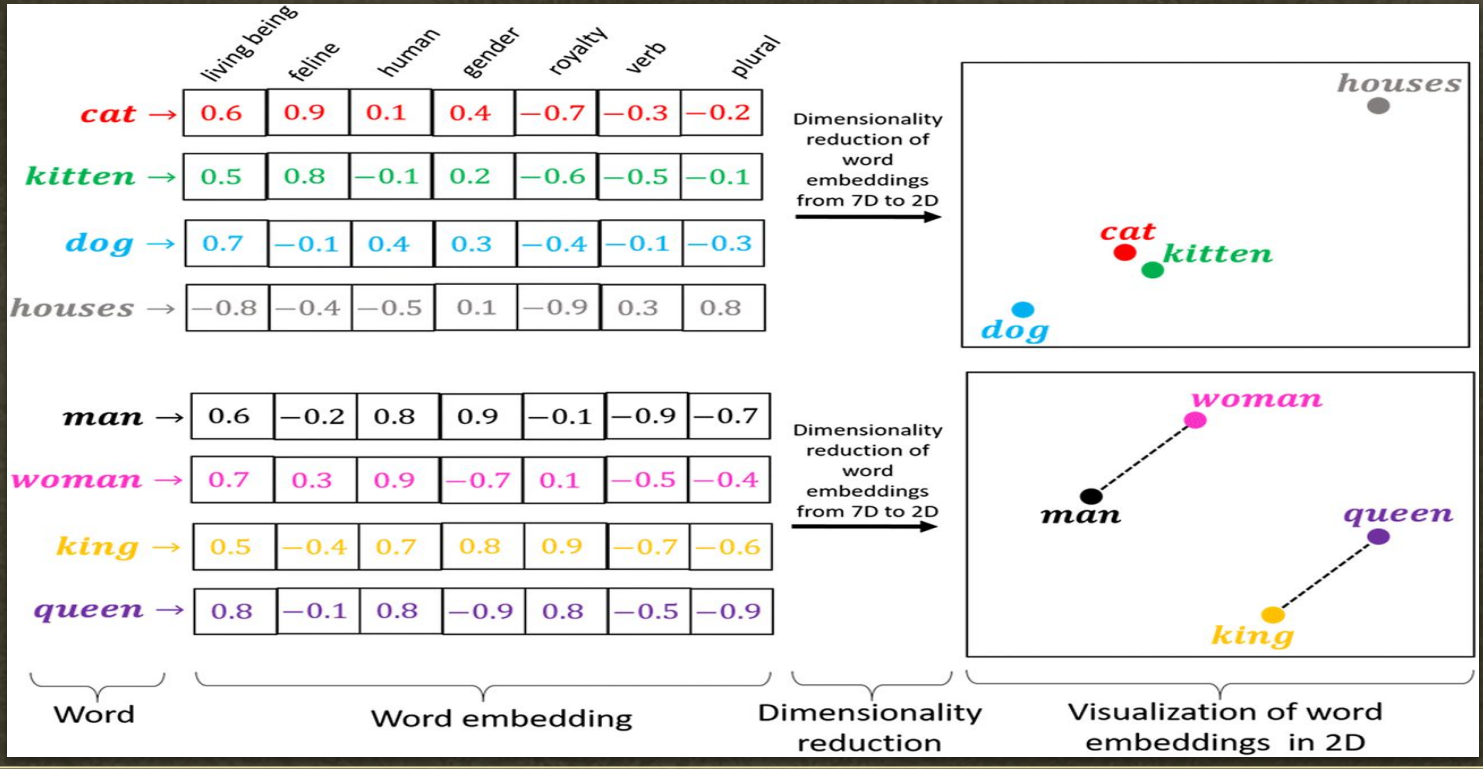
Resultados

Nivel de ruido BCR	Blue score
low	0.5300
medium	0.4355
high	0.3591

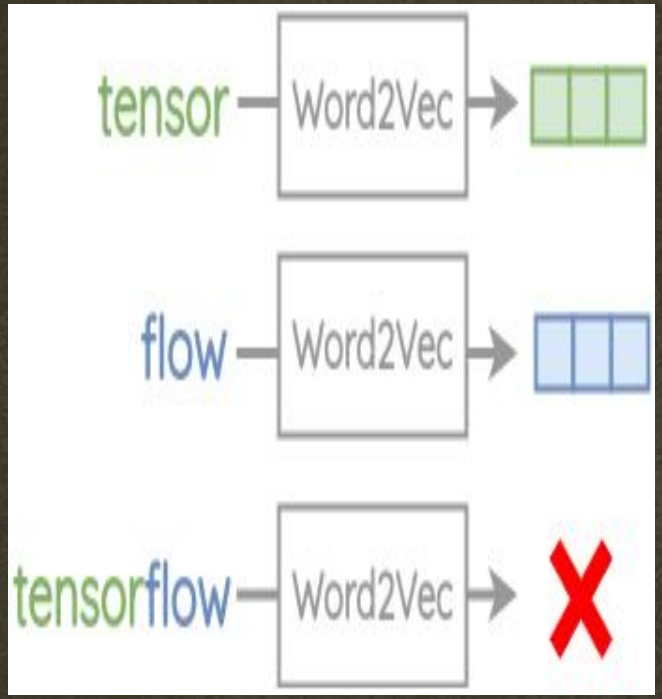
Método 2

método basado en word embeddings

Embeddings


living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

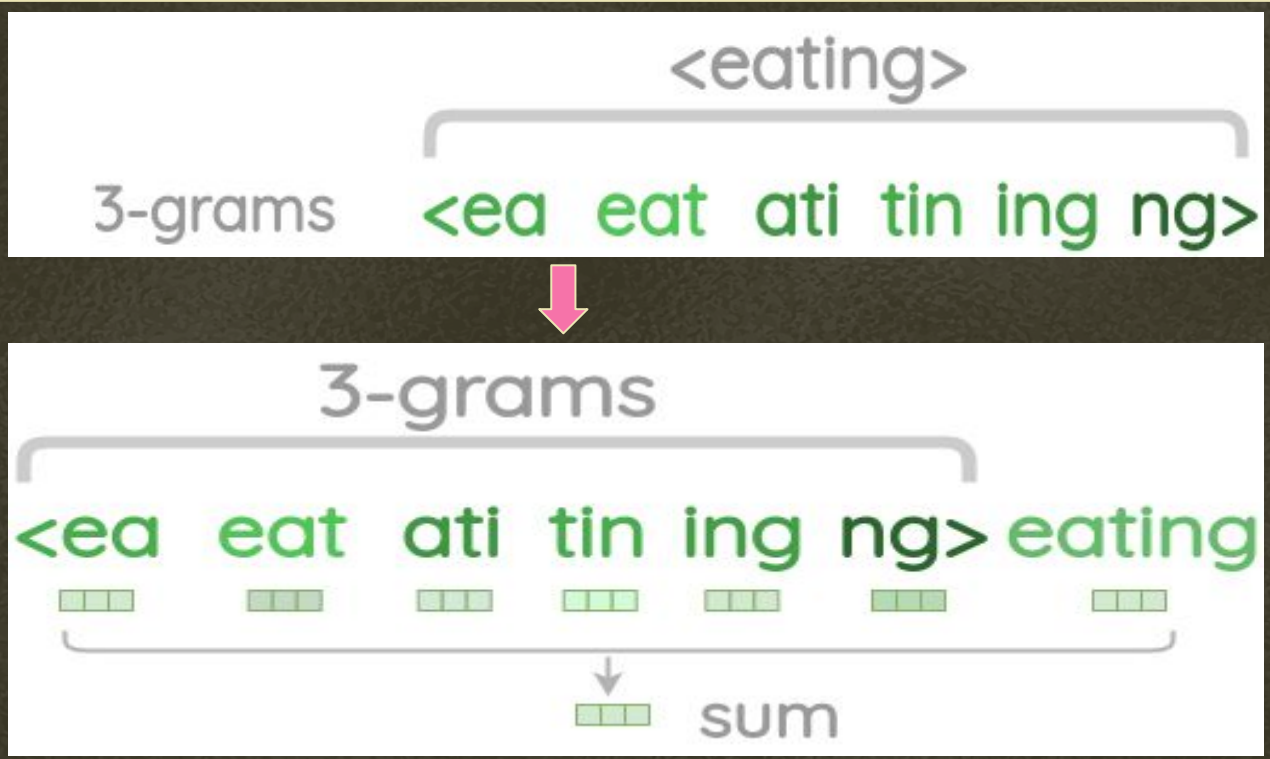
Problemas wor2vec

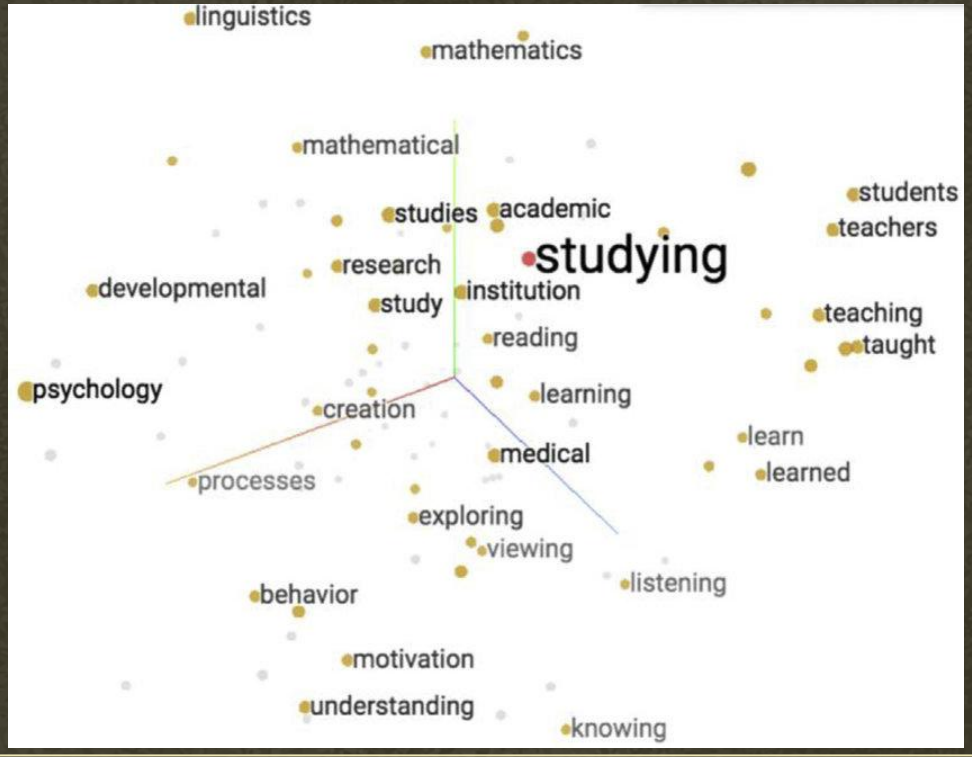


Shared radical

eat eats eaten eater eating

Fasttext


<eating>
3-grams
<ea eat ati tin ing ng>
3-grams
<ea eat ati tin ing ng> eating
sum

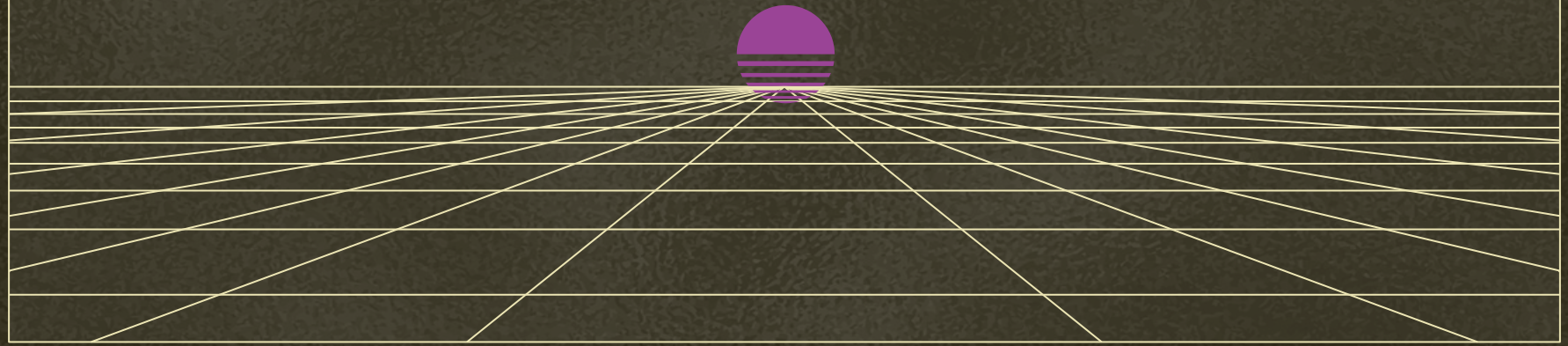

linguistics
mathematics
mathematical
students
studies
academic
teachers
studying
research
developmental
institution
study
teaching
reading
taught
psychology
learning
creation
learn
medical
learned
processes
exploring
viewing
listening
behavior
motivation
understanding
knowing

ORACION ORIGINAL
gobernadores, todos del mismo paktid0, los pkesidente8

ORACION CORREGIDA
gobernadores todos del mismo paktiyā los presidente de

ORACION ORIGINAL
pke8idente utilizaba. Bn otro nivel de la pikámide estaban lus

ORACION CORREGIDA
idente utilizaba plutonio otro nivel de la pirámide estaban

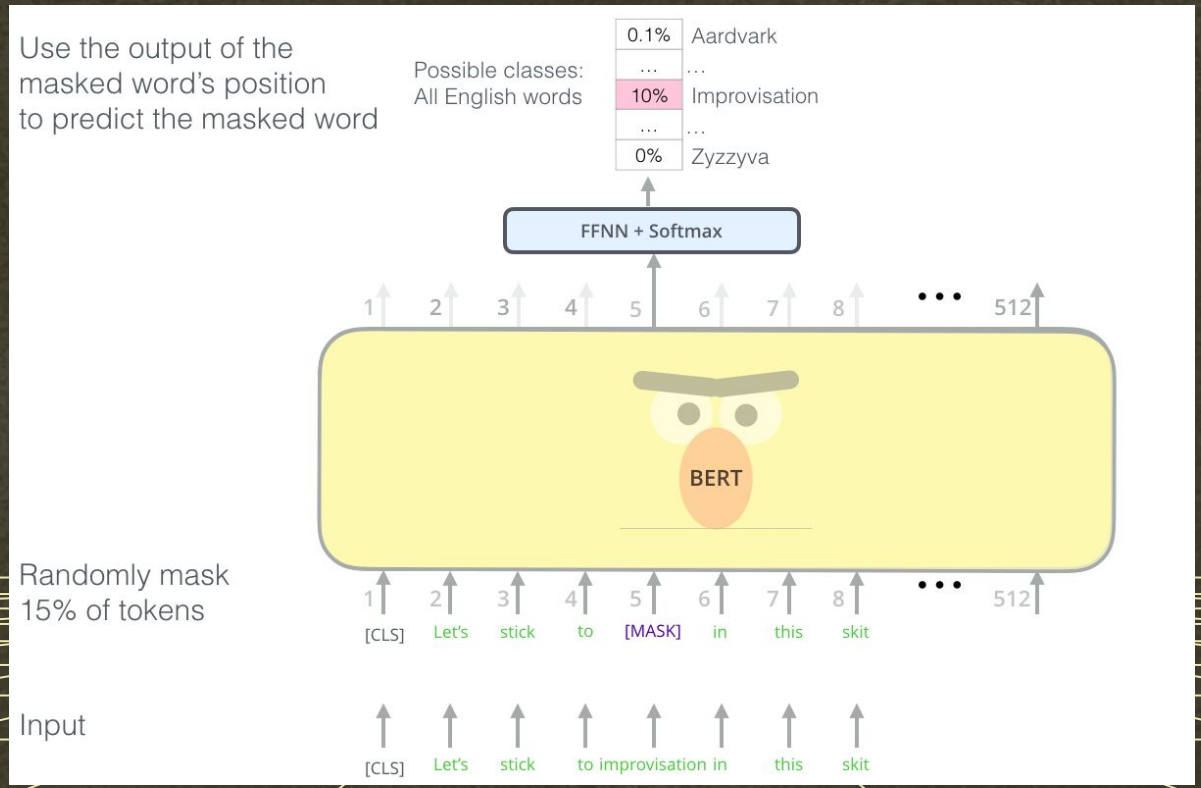


Resultados con Fasttext

Nivel de ruido OCR	Blue score
low	0.5025
medium	0.4162
high	0.3034

Método 3

método basado en transformers

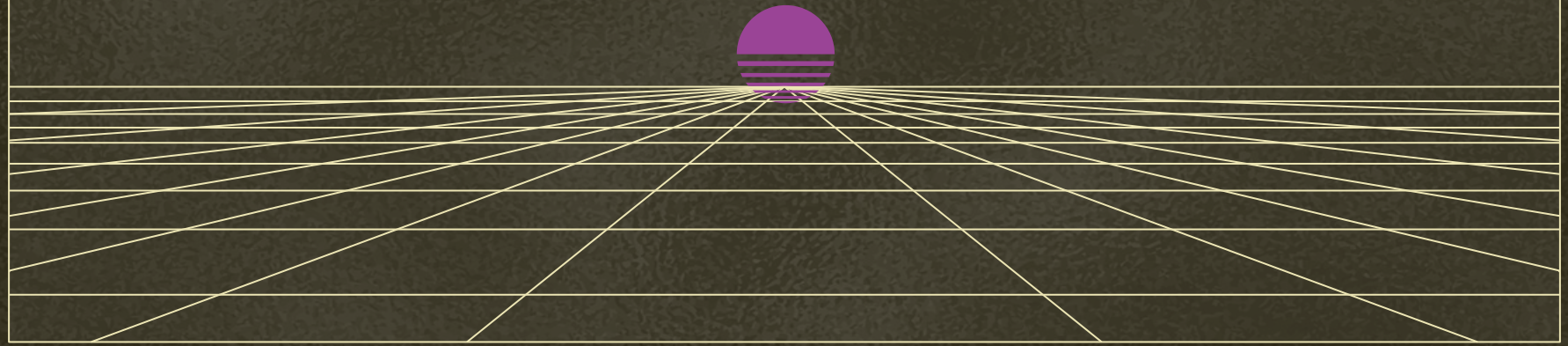

Use the output of the masked word's position to predict the masked word
Possible classes: All English words
0.1% Aardvark
... ...
10% Improvisation
... ...
0% Zyzzyva
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
Randomly mask 15% of tokens
1 2 3 4 5 6 7 8 ... 512
[CLS] Let's stick to [MASK] in this skit
Input
[CLS] Let's stick to improvisation in this skit

ORACION ORIGINAL
setenta han pu61icadu obras testimoniales en on e8foekzo por recuperar la memoria

ORACION CORREGIDA
setenta han publicado obras testimoniales en un desarrollo por recuperar la memoria

ORACION ORIGINAL
histórica de aquellos años. Sin em6ar90, esos testimonios abarcan en 8u mayoría a la

ORACION CORREGIDA
histórica de aquellos años . sin embargo , esos testimonios abarca en su mayoría a la



Resultados con BETO

Nivel de ruido ECR	Blue score
low	0.67899
medium	0.6421
high	0.5092

Siguientes pasos

Vocabulario

Formar un vocabulario más grande

Preprocesamiento

Mejorar el preprocesamiento del texto

Neuspell

Hacer funcionar neuspell para nuestro propósito

Gracias