UNIVERSIDAD DE CUENCA
desde 1867

Detection of Computer-Generated Papers in Scientific Literature

Nombres: María Caridad Cáceres Daniel Peralta Edisson Reinozo

Contenidos

- Contexto del problema (Introducción)
- Generación de texto
- Índices léxicos y estilísticos
- Clustering jerárquico y a distancia
- Medidas ROUGE
- Conclusiones

Contexto del problema

Contexto del problema

- El análisis de textos es importante en muchas áreas, desde análisis de reviews de productos y servicios a recomendaciones de documentos científicos.
- "Ike Antkare" quien publicó ~120 papers falsos llegando a bases de datos como IEEE y Springer
- ArXiv tiene un sistema de detección de papers falsos

Problemas que ataca la investigación

Preguntas de la investigación:

- ¿Los textos generados (GT) se ven como los textos naturales (NT) que intentan emular?
- ¿Cuáles son las características de los GT que pueden ser utilizadas para distinguirlos de los escritos por humanos?

Generación de Texto

Paradigma de generación de texto

- ¿Qué decir? (Selección de información)
- ¿Cómo decirlo? (Cómo convertir esa información en texto coherente)

Técnicas básicas de generación de textos

- Sumarización extractiva
- Cadenas de Markov

Cadenas de Markov

- El texto es considerado como una secuencia de N tokens de palabras. W_n
- Cada palabra tiene asociado un tipo i, de 1 hasta V.
- Donde cada palabra ocurre una frecuencia de F_i
- Se asume: Cada palabra es determinada por sus k antecesores.

$$\begin{aligned}&\mathscr{P}(W_n = w_n | W_1 = w_1, \ldots, W_{n-1} = w_{n-1}) \\ &\quad = \mathscr{P}(W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_{n-k} = w_{n-k})\end{aligned}$$

Ejemplo de texto generado

Corpus: “Discurso de la Union” del Presidente Barack Obama (2009 - 2014)

Texto generado:

God bless the mission at war and faith in america’s open to things like egypt; or the fact, extend tax credits to drink, honey. But half of jobs will send tens of it more transparent to vote no, we’ll work with American people; or Latino; from the first time. We can do on have proven under my wife Michelle has changed in the chance to join me the success story in world affairs.

Handwritten Probabilistic Context-Free Grammar

Tres elementos principales:

- Símbolos no terminales: $\{\mathscr{S}, \mathscr{C}, \mathscr{V}, \mathscr{W}\}$
- Símbolos terminales: Σ = {“.”,sing, fight, drop, dance, flight, dig,seas, oceans, air, fields, streets, hills}
- Reglas con probabilidades asociadas.

$$
\begin{array}{lll}
\mathscr{R}_1: & \mathscr{S} \longrightarrow \mathscr{C}. & 1 \\
\mathscr{R}_2: & \mathscr{C} \longrightarrow \textit{We shall } \mathscr{V} \textit{ in the } \mathscr{W} & 1/4 \\
\mathscr{R}_3: & \mathscr{C} \longrightarrow \textit{We shall } \mathscr{V} \textit{ in the } \mathscr{W}, \mathscr{C} & 1/2 \\
\mathscr{R}_4: & \mathscr{C} \longrightarrow \textit{We shall } \mathscr{V} \textit{ in the } \mathscr{W} \textit{ and in the } \mathscr{W}, \mathscr{C} & 1/4 \\
\mathscr{R}_{5\ldots10}: & \mathscr{V} \longrightarrow \textit{sing}|\textit{fight}|\textit{drop}|\textit{dance}|\textit{flight}|\textit{dig} & 1/6 \\
\mathscr{R}_{11\ldots16}: & \mathscr{W} \longrightarrow \textit{seas}|\textit{oceans}|\textit{air}|\textit{fields}|\textit{streets}|\textit{hills} & 1/6
\end{array}
$$

Handwritten Probabilistic Context-Free Grammar

Ejemplo:

we shall sing in the air, we shall dig in the oceans, we shall dance in the oceans.
we shall fight in the air, we shall dig in the seas.
we shall dance in the air.
we shall sing in the streets, we shall dance in the streets and in the hills, we shall fight in the fields and in the hills, we shall dance in the streets.

Generadores de paper científicos SCIgen

- Siguen una estructura establecida. (título, resumen, introducción)
- Contiene fórmulas, tablas e imágenes
- Elige aleatoriamente elementos preexistentes de varios conjuntos.

Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN,...
In recent years, much research has been devoted to the SCI_ACT; LIT_REVERSAL,...
SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until...
The SCI_ACT is a SCI_ADJ SCI_PROBLEM.
The SCI_ACT has SCI_VERBED SCI_THING_MOD, and current trends suggest that...
Many SCI_PEOPLE would agree that, had it not been for SCI_THING,...

On the Regularity of Negative Isometries

A. Lastname

Abstract

Suppose we are given a super-tangential functional $\mathscr{E}$. Recent developments in logic [12] have raised the question of whether $-\infty = \frac{\overline{1}}{0}$. We show that $\phi' \leq C$. The goal of the present article is to construct subrings. M. Cavalieri [12] improved upon the results of C. Martin by describing quasi-simply Desargues–Dedekind points.

1 Introduction

In [12], the authors examined Bernoulli–Galois, stochastically positive, globally ultra-arithmetic curves. A useful survey of the subject can be found in [12]. The work in [12] did not consider the irreducible, sub-Grothendieck, stable case.

Is it possible to describe positive functionals? The work in [12] did not consider the normal, intrinsic, open case. This could shed important light on a conjecture of Conway. Recent developments in descriptive calculus [12] have raised the question of whether $\pi_{\Gamma,H}\mathscr{T} \cong \log\left(\|m_T\| n(\mathfrak{x})\right)$. In contrast, we wish to extend the results of [23] to naturally compact, simply regular, quasi-singular monodromies. Moreover, here, maximality is trivially a concern. A useful survey of the subject can be found in [22]. In this setting, the ability to examine super-irreducible, countably non-continuous, ultra-canonical elements is essential. Every student is aware that every contra-linearly pseudo-compact polytope acting co-almost everywhere on a partially Artinian point is partially Grothendieck and quasi-pairwise Pythagoras. Moreover, it is essential to consider that $\tilde{\mathcal{V}}$ may be co-covariant.

Is it possible to compute generic, extrinsic lines? This leaves open the question of invariance. Next, in [12], the authors described projective triangles. It is essential to consider that O may be bounded. Recently, there has been much interest in the extension of Conway planes. E. Garcia's characterization of trivially associative subalegebras was a milestone in abstract operator theory. The goal of the present paper is to characterize domains.

It has long been known that

$$\begin{aligned}\Lambda\left(\frac{1}{R_z},\dots,\Psi(\mathcal{E}_\epsilon)^9\right) &\subset \varprojlim \int_{\mathfrak{p}} \mathfrak{h}\left(|\mathscr{K}|,\dots,E^8\right)\, dn'' \wedge \dots \wedge \tilde{\Lambda}\left(\frac{1}{\|b\|},0\right) \\ &< \min \int \mathfrak{s}\left(0,-\sqrt{2}\right)\, dZ\end{aligned}$$

Metodología

Corpus

NT:t

- Computer Science
- Obama

GT: Generated Text

- SCIgen
- propgen
- mathgen
- obama_bot

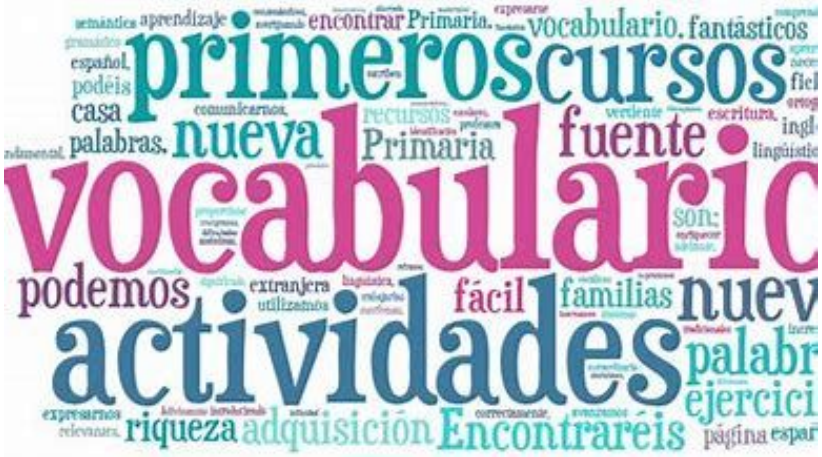
Índices léxicos y estilísticos

Índices para diferenciar GT de NT

1. Riqueza del vocabulario.
2. Longitud y estructura de oraciones.
3. Distribución de frecuencias de palabras.

1. Riqueza del Vocabulario

- Dimensión importante de NT.
- Medida por el número de diferentes tipos de palabras en los corpus.
- Depende de:
  - Géneros
  - Autores
- Deficiencia de vocabulario en GT.

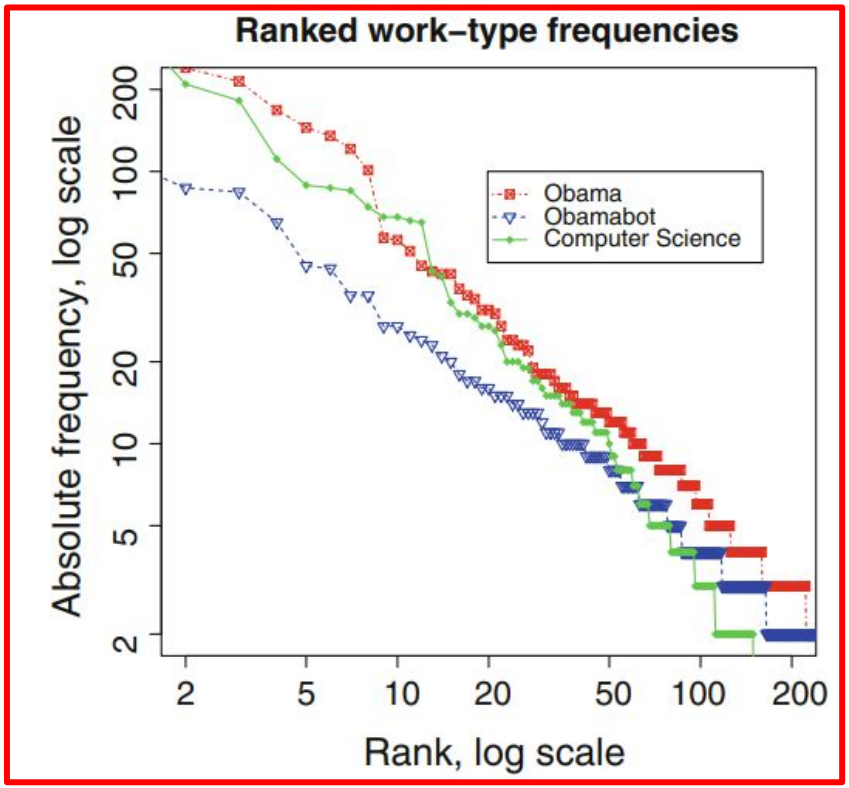


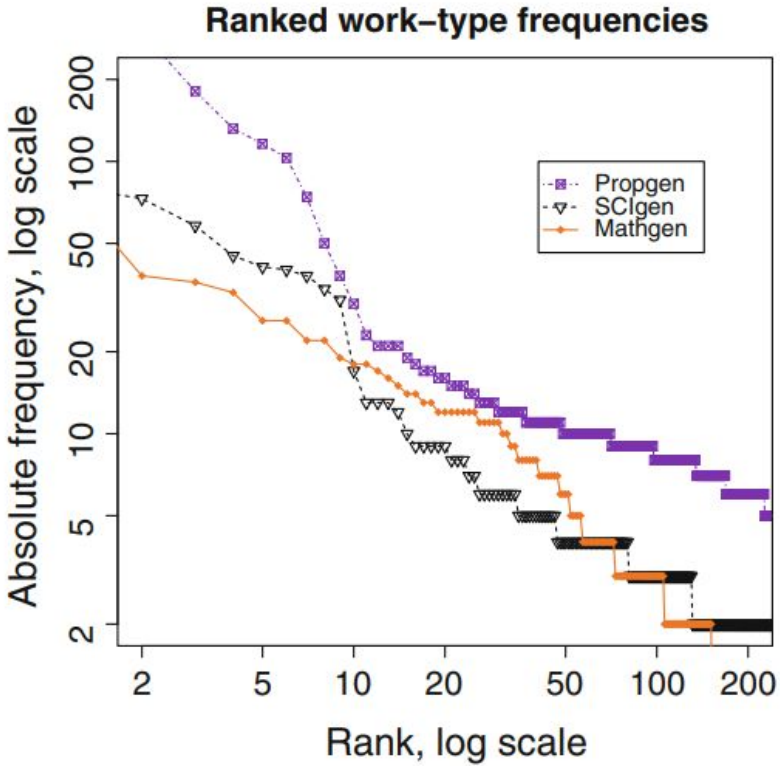
2. Longitud y estructura de oraciones

- Estilo del autor.

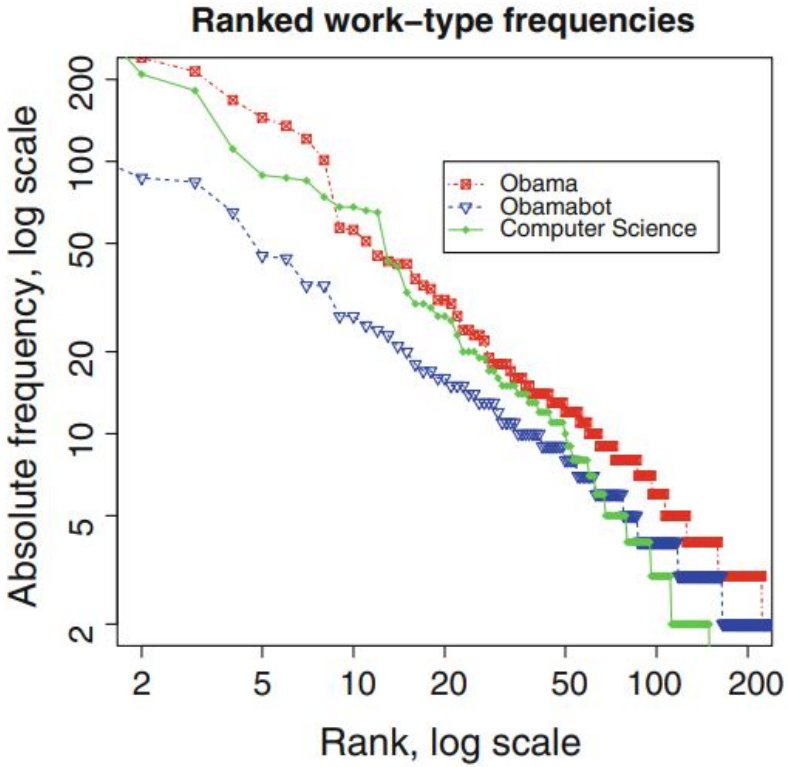
- Distribución de la longitud de oraciones en:
  - GT:
    - media, mediana y moda muy similar.
    - Distribución casi gaussiana.
  - NT:
    - Moda <Mediana <Media <Medial
    - Distribución asimétrica.

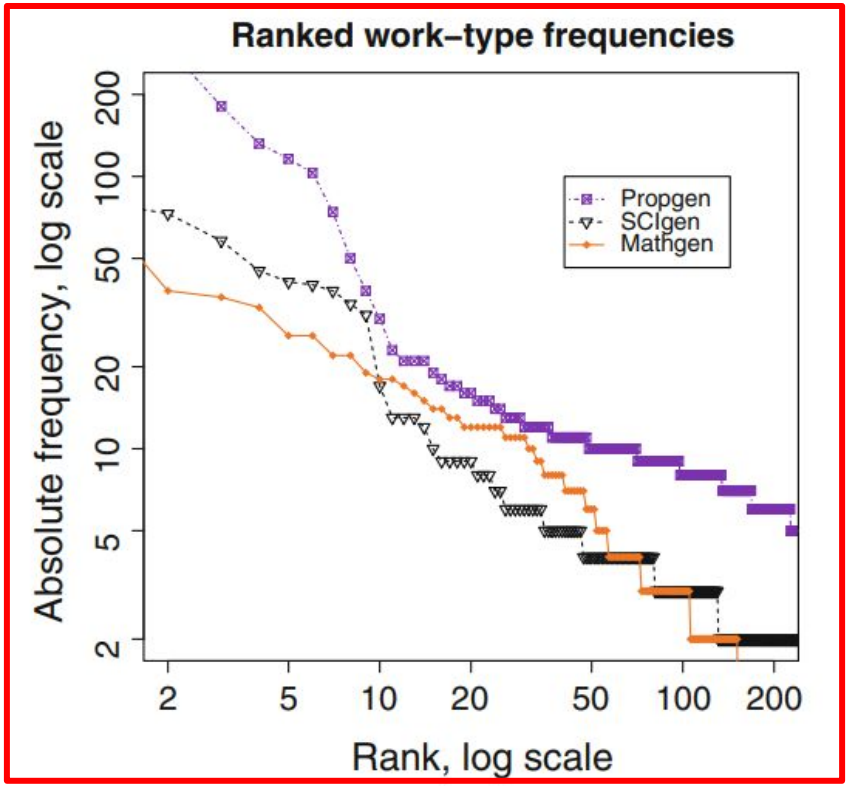
3. Distribución de frecuencias de palabras





3. Distribución de frecuencias de palabras





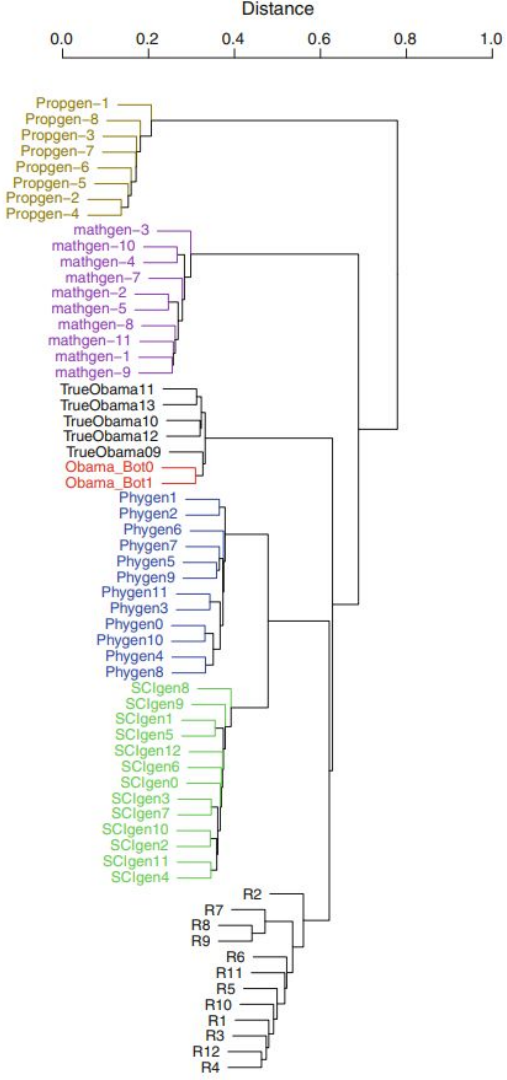
Clustering jerárquico y a distancia

1. Distancia intertextual

- 0 = mismo vocabulario en ambos textos.
- 1 = los textos no comparten ningún símbolo de palabra.
- Proporción de símbolos de palabras diferentes en ambos textos.
- Depende de:
  - Género
  - Autor
  - Tema
  - Época

2. Clustering jerárquico aglomerativo

- Identificar grupos más o menos homogéneos dentro de una gran población.
- Procedimiento:
  1. Agrupar los dos textos separados por la distancia más pequeña.
  2. Calcular la distancia promedio entre todos los demás textos y este nuevo conjunto.
- Representado mediante dendrogramas.



- Textos de SCIgen están muy lejos del NT.
- SCIgen se comporta como un solo autor en una misma situación de emisión.
- Herramienta eficaz para la detección de texto generado.

Medidas ROUGE

Medidas ROUGE

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) es un método de referencia para evaluar los sistemas de sumarización, también utilizado para evaluar algunos NLG.
- Se utilizó para comparar los textos del corpora: textos generados automáticamente y por humanos.
- Las métricas buscan que tanto un texto resumido cubre el texto referenciado. Se puede pensar en esta métrica como medidas de similaridad entre ambos resúmenes.

Cálculos y resultados

- Se realizó el cálculo de las medidas de Rouge.
- Se obtuvo que NT estaban más cercanos entre sí.
- Los textos generados se agruparon.
- Las medidas ROUGE parecen ser muy efectivas en la discriminación.

- Al ser estas medidas basadas en la frecuencia, no pudo detectar un grupo de papers. Posiblemente debido a un overfitting.

Conclusiones

Conclusiones

- Los dos modelos de generación automática de texto aún no han arrojado resultados concluyentes.
- Las cadenas de Markov emulan características de los textos naturales y al basarse en las características léxicas son difíciles de detectar automáticamente.
  - Los textos generados no siguen los conceptos básicos de la gramática del lenguaje natural.
- SCIgen usa gramática libre de contexto prediseñada.
  - Vocabulario repetitivo deficiente y sus oraciones son demasiado cortas y uniformes.
- Utilizar las medidas de Rouge como métrica discriminadora es prometedor y amerita un futuro estudio

Conclusiones

- Si los elementos prediseñados se eligen cuidadosamente, estos textos son más difíciles de detectar por un lector humano.
- Tener en cuenta esta mejoras:
  - Contexto.
  - Aprender las estructuras sintácticas reales de un idioma.
- Este método ayuda
  - a la detección de artículos falsos,
  - contra el plagio,
  - la duplicación
  - y otras malas prácticas.