Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Curso: Análisis de Algoritmos
Profesor: José Carranza-Rojas

Semestre II, 2017

Valor: 15%
Proyecto en parejas

Proyecto 3 - El Poeta

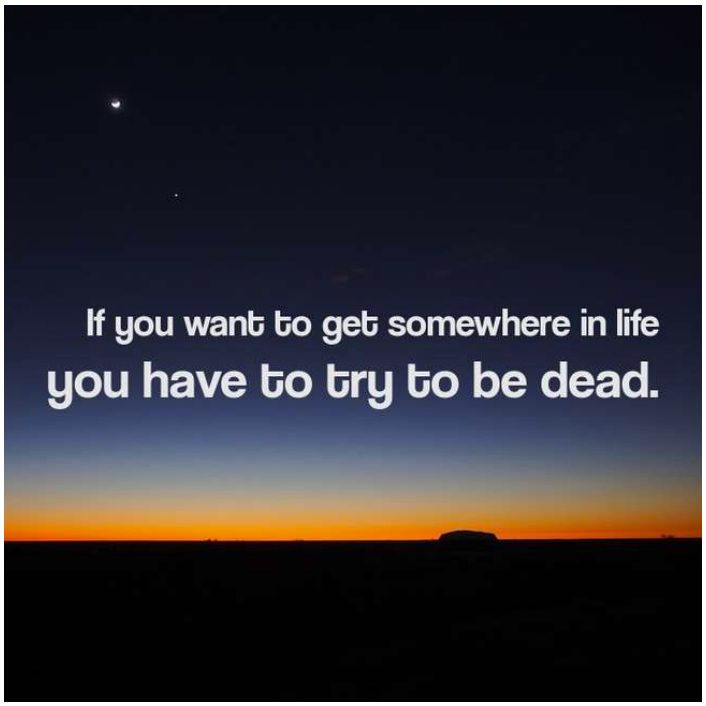

Figure 1 – Frases inspiradoras autogeneradas con IA, http://inspirobot.me/

La generación automática de texto es un problema aún abierto en inteligencia artificial y procesamiento de lenguaje natural. En este proyecto crearemos un algoritmo genético que, usando distancias de similitud entre documentos (poemas en inglés) como función de adaptabilidad, estará en la capacidad de aproximar un poema meta dado por el usuario, para crear poemas automáticamente, que se parezcan al poema meta.

Claramente, las técnicas a usarse en el proyecto no garantizan que necesariamente haya una buena semántica en los textos resultantes. En otras palabras, puede que los textos generados sean un tanto… ridículos.

Diccionario y n-grams

Esta técnica se usa para convertir un documento o texto en una serie de tuplas con la intención de capturar las interacciones entre las diferentes palabras del texto. La idea es muy simple: se guarda en un diccionario o vector las n palabras continuas pegadas como un item individual, y se puede variar el n. Por ejemplo, cuando n=1, entonces cada palabra se guarda por separado (no hay interacción). Si n=2, entonces se forman tuplas de 2 palabras continuas, si n=3 serían 3 palabras continuas que aparecen en el texto, y así sucesivalemtne. Si se calculan todos los n-grams para un conjunto de documentos, se puede crear un diccionario de n-grams donde cada n-gram tiene su propio codigo, permitiendo convertir el texto en un vector de números, que eventualmente puede ser comparable con otro.

Cálculo de similitud entre documentos

Una vez calculados los n-grams, un documento puede ser convertido en un histograma que revele la frecuencia con que aparece cada n-gram. Por ejemplo, si se hace un hsirograma de 1-grams, revelaría cuantas veces aparece cada palabra sola del diccionario. Con esta nueva representación, es posible calcular índices de similitud para saber qué tanto se parecen dos documentos, ya que comparar dos documentos se vuelve un problema más sencillo que sería comparar dos histogramas.

Se deben implementar 4 distancias entre histogramas de documentos:

Instituto Tecnológico de Costa Rica Semestre II, 2017
Escuela de Ingeniería en Computación
Curso: Análisis de Algoritmos
Profesor: José Carranza-Rojas
Valor: 15%
Proyecto en parejas

1. Manhattan
2. Chebyshev
3. Distancia Propia: Deben inventarse una distancia propia y usarla en los experimentos, que funcione al menos tan bien como la Manhattan.

Algoritmos Genéticos

La forma de generar los poemas automáticamente será con algoritmos genéticos, con el fin de aproximar otro poema que se usará como poema meta. Una vez tomados todos los textos de los poemas del set de datos y que se ha crreado el diccionario de n-grams (para varios n), se procederá a generar aleatoriamente la población inicial de documentos, tomando un número de de n-grams aleatorios y concatenandolos. A partir de esta pobleción inicial, podremos escoger los documentos más aptos, calculando con algun índice de similitud la distancia que hay entre el poema meta, y cada individuo. Esto permitirá la creación de la siguiente población, así como de mutaciones y cruces.

En cuánto a las mutaciones y cruces, deben describir muy bien en el documento de Latex cómo se hicieron y quedan a criterio abierto de los estudiantes, siempre buscando maximizar la obtención de los mejores resultados posibles. Recordar tener en cuenta evitar que hayan poblaciones degeneradas.

La idea es que despues de muchas generaciones, los individuos (documentos, textos, poemas) más aptos, sean la respuesta al problema, y sean poemas parecidos al poema meta.

Lenguaje de Programación e Interfaz

Se debe programar en C# con Visual Studio. La aplicación debe permitir en la interfaz escoger un archivo de texto con el poema meta (o escribirlo en un text field), y mostrar la lista de los k poemas autogenerados más aptos, despues de un número dado de generaciones. También debe permitir escoger el método de similitud a usarse.

Datos

Usaremos un set de datos dado por el sitio web Kraggle:

https://www.kaggle.com/ultrajack/modern-renaissance-poetry

Documento en Latex / PDF

Deben crear en latex (y entregar los fuentes de latex) un documento donde describen el análisis de cada distancia de similitud, incluyendo la propia. Además deben incluir experimentos donde se compara el rendimiento de las diferentes distancias de similitud, así como el uso de n-grams. Cuál es un buen n para comparar poemas? Habrá un n que ya no aporta en la obtención de buenos resultados en la función de adaptabilidad? Habrá una relación entre el n y el tipo de función de similitud? Cuántas generaciones se necesitan para obtener poemas con distancias bajas?

Instituto Tecnológico de Costa Rica Semestre II, 2017
Escuela de Ingeniería en Computación
Curso: Análisis de Algoritmos
Profesor: José Carranza-Rojas
Valor: 15%
Proyecto en parejas

Se espera que hagan al menos 3 experimentos para responder preguntas como estas, y que describan los resultados y los ***analicen***.

El documento debe seguir el template de la IEEE para publicaciones en ingeniería el cual pueden encontrar aquí:

https://www.ieee.org/documents/ieee-latex-conference-template.zip

Evaluación

Tarea	Puntaje Máximo
Programación completa de genéticos	20
Distancia Manhattan (Programación)	5
Distancia Manhattan (Análisis)	5
Distancia Chebyshev (Programación)	5
Distancia Chebyshev (Análisis)	5
Distancia Propia (Diseño y Programación)	10
Distancia Propia (Análisis)	5
Calculo de n-grams e histogramas	10
Experimentos y discusión (3 experimentos)	30
Interfaz	5

Figure 2– Frases inspiradoras autogeneradas con IA, http://inspirobot.me/