Detección de *bots* en Twitter*

18 de julio de 2018

Descripción del proyecto: En Twitter se producen alrededor de 6.000 tweets por segundo. Las publicaciones de Twitter son en su mayoría públicas y pueden recopilarse fácilmente mediante una API.

El papel de los llamados "*bots*", que son cuentas automatizadas capaces de publicar contenido o interactuar con otros usuarios, ha sido objeto de discusión en los últimos años [1]. Estas cuentas pueden desempeñar un papel valioso en las redes sociales, por ejemplo, si son usadas para responder preguntas sobre variedades de temas en tiempo real, o para proporcionar actualizaciones automáticas sobre noticias o eventos. Sin embargo, los *bots* también se pueden usar para difundir información errónea (*fake news*), alterar percepciones políticas, o manipular los sistemas de clasificación y revisión en línea [2]. A medida que las redes sociales han ido alcanzando una posición prominente como medios de comunicación masivos, el control de los *bots* se ha vuelto una tarea relevante [3].

Para detectar *bots*, se pueden incorporar modelos de clasificación y/o técnicas de procesamiento del lenguaje natural, tales como *topic modeling* y *sentimental analysis*. Este proyecto implicará el uso de ingeniería de factores y a su vez requerirá la recopilación de datos del mundo real.

Objetivo general: emplear técnicas de Machine Learning para detectar *bots* utilizando datos provenientes de tweets recopilados con la API de Twitter.

Bases de datos: para este proyecto, los datos deben recolectarse usando la API de Twitter. Tendrán a disposición un *script* de Python, `tweepy_srcipt.ipynb`, que utiliza la librería *tweepy* [4] para acceder a la API de Twitter. El archivo de ejemplo `tweets_sample.json` es una muestra del tipo de datos a ser recolectados. Además, este archivo muestra los atributos de un tweet que están disponibles al usar la API [5].

Objetivos específicos: .

1. El primer paso es crear su propia base de datos, use la API de Twitter, ingeniería de factores y técnicas de pre-procesamiento para preparar sus datos para el análisis posterior.

*Proyecto formulado por el Profesor Pavlos Protopapas

2. Cree modelos para determinar las características de los diferentes tipos de usuarios de Twitter. Se recomienda crear al menos un modelo que use técnicas de procesamiento del lenguaje natural, por ejemplo *topic modeling* [6], y al menos un modelo que use un algoritmo de clasificación. También puede usar modelos que usen ambas técnicas.
3. Proporcione evidencia de éxito del modelo(s) propuesto(s) para la detección de *bots*, muestre también las limitaciones de su(s) modelo(s). Sería recomendable que incluya un análisis de error y una evaluación de la calidad de predicción que tienen su(s) modelo(s).

Referencias: .

1. Stefan Wojcik, "Bots in the Twittersphere": http://www.pewinternet.org/2018/04/09/bots-in- the-twittersphere/
2. Chris Baraniuk, "How Twitter Bots Help Fuel Political Feuds": https://www.scientificamerican.com/article/how-twitter-bots-help-fuel-political-feuds/
3. Chengcheng Shao et al., "The spread of low-credibility content by social bots": https://arxiv.org/pdf/1707.07592.pdf
4. The tweepy Python library: http://www.tweepy.org
5. Twitter's developer resources to learn about the API: https://developer.twitter.com
6. Asbjørn Ottesen Steinskog et al., "Twitter Topic Modeling by Tweet Aggregation": http://www.aclweb.org/anthology/W17-0210