Using Text Mining and Clustering Techniques on Tweets to Discover Trending Topics in Dubai

استخدام تقنيات استنباط النصوص على تغريدات تويتر لاكتشاف أكثر المواضيع تداولا عن دبي

By
Moutaz Wajih Hamadeh
ID#: 120032

Dissertation submitted in partial fulfillment of
Msc in Information Technology Management

Faculty of Engineering & Information Technology

Dissertation Supervisor
Dr. Sherief Abdullah

May-2015

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Moutaz Hamadeh	120032	IT Management	31 May 2015

Title
Using Text Mining and Clustering Techniques on Tweets to Discover Trending Topics in Dubai

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make that copy available in digital format if appropriate.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature

Abstract

Twitter micro-blogging website is a hot and emerging area of research recently, users on Twitter post millions of tweets every day from all over the world, it is very difficult and challenging task to keep track of messages and filter them based on topical interest. This study uses text mining and clustering techniques to partition Dubai tweets into clusters of a same topical interest. Tweets corpus of Dubai were collected, they were presented through the bag of words model using TF-IDF weighting scheme, after that the output of text transformation was introduced to k-means clustering algorithm with cosine similarity measure. Findings indicate that heuristic evaluation techniques are not so helpful in this domain; also, the model has generated interesting clusters about trending topics and events in Dubai. In the end, an experiment was conducted over datasets collected from different timeframes to see what are the constant hot topics discussed in Twitter about Dubai. All of the findings in this report have been empirically analysed through real-word tweets dataset.

الخلاصة

يعتبر نظام المدونات المصغرة "تويتر" أحد أبرز النطاقات سخونة و خصوبة في مجال البحث العلمي, حيث يقوم يوميا مستخدمو النظام بنشر عدة ملايين من التغريدات من جميع أنحاء العالم. و للحجم الضخم جدا من المعلومات في تويتر, فإن عملية مراجعة جميع التغريدات و تنقيتها بناء على مواضيعها تعتبر من التحديات الصعبة جدا. هذه الدراسة تستخدم تقنيات استنباط النصوص لتقسيم التغريدات التي تخص مدينة دبي إلى عدة مجموعات, كل مجموعة تهتم بموضوع معين. من أجل ذلك, تم تجميع قاعدة من التغريدات التي تخص دبي و عرضها باستخدام نظام عرض النصوص TF-IDF

النتيجة التي تم الحصول عليها من نظام العرض TF-IDF, تم تقديمها إلى خوارزمية التجميع المعروفة بـ K-means. النتائج التي تم الحصول عليها تشير إلى أن أساليب التقييم الارشادية غير مفيدة للتطبيق على هذا النطاق من الدراسة. كما أن نموذج الدراسة أنتج عدد من المجموعات عن المواضيع الساخنة في دبي. و في نهاية الدراسة, تم إجراء اختبار بتجميع عدد من التغريدات على عدة مراحل زمنية من أجل معرفة أكثر المواضيع الدائم تداولها عن دبي عبر تويتر. جميع الدراسات التي طبقت في هذه الأطروحة تم تحليلها بناء على قاعدة بينات حقيقية تم استنباطها من تويتر.