# Giới thiệu bài toán - Dũng

Mục đích

Dữ liệu đầu vào

Các bước giải quyết vấn đề (mô tả dữ liệu, kết xuất feature, phân loại bằng mô hình ML, phân loại bằng mô hình DL, độ đo và đánh giá).

Công cụ pyspark

# Mô tả dữ liệu - Dũng

# Trích chọn đặc trưng (một số phương pháp) – Hồng

Trường “text” chứa câu tweet nguyên bản chứa 1.350.484 từ khác nhau. So với số lượng câu tweet trong dataset là 1.600.000 câu thì số lượng từ như vậy là khá lớn. Vì vậy, ta phải tiền xử lý dữ liệu trước khi trích xuất đặc trưng từ trường text. Mục đích là chuẩn hóa để giảm số từ trong từ điển xuống mức thấp so với số câu thì mô hình phân loại mới hoạt động tốt được.

Để phân tích cảm xúc từ câu tweet ta xử lý lần lượt qua 1 số bước sau:

* Tiền xử lý trường “text” chứa câu tweet:
  + Dùng Regular Expression để xóa các từ bắt đầu bởi “@...” (ví dụ @switchfoot). Lý do là vì đây là tên riêng của chủ tài khoản Twitter được nhắc đến. Đây là cách giao tiếp giữa các tài khoản Twitter, nên dạng từ như này khá phổ biến. Tuy nhiên, vì là tên riêng nên nó sẽ không giúp ích gì cho việc phân loại cảm xúc của câu tweet mà chỉ gây nhiễu cho mô hình phân loại.
  + Dùng Regular Expression xóa các từ bắt đầu bởi “http…”. Số lượng từ như này ít hơn so với từ “@...” nhưng cũng tương tự như từ “@...”, nó không giúp ích gì cho việc nhận biết cảm xúc của câu.
  + Loại bỏ những câu có độ dài không quá 10 kí tự, vì câu quá ngắn nên khó nhận biết cảm xúc với ít nội dung như vậy.
  + Biến câu tweet thành chuỗi token bằng thàm RegexTokenizer với tùy chọn minTokenLength = 1 để loại bỏ các token rỗng (“”) là kết quả tokenize từ các dấu phân cách.
  + Loại bỏ các từ dừng, là các từ phổ biến trong tiếng Anh, vốn không giúp ích trong việc nhận diện cảm xúc của câu, dùng StopWordsRemover().
  + Đưa các từ trong câu tweet về dạng từ đơn giản, loại bỏ các tiền tố và hậu tố bằng thư viện nltk.stem.snowball. (thử đưa về lemma)
* Tiền xử lý trường “target” chứa nhãn. Trường này chứa 2 giá trị dạng số là 0 và 4 (0 = negative, 4 = positive). Ta phải đưa nó về dạng thứ bậc (index) để làm nhãn cho mô hình phân loại.
* Tạo thêm trường khung giờ từ trường “date”
* Tạo thêm trường độ dài câu tweet
* Trích xuất đặc trưng từ nội dung câu tweet sau khi đã chuẩn hóa. Ta thử 1 số phương pháp sau:
  + TF-IDF (với CountVectorizer)
  + TF-IDF (với HashTF)
  + Word2Vec
  + N-gram

# Phân loại bằng mô hình ML (so sánh và lựa chọn) – Hồng

# Phân loại bằng mô hình DL (so sánh và lựa chọn) - Hưng

# Kết luận và hướng phát triển - Hưng

Kỹ thuật streaming để lấy câu tweet online và có kết quả realtime để theo dõi quan điểm của cộng đồng về 1 chủ đề cụ thể nào đó (ví dụ bitcoin, mã chứng khoán…). Việc này sẽ rất có ích cho các quyết định mua, bán