Thesis: Text Mining on reported events

Context

With the development of digitalization in air transport, the amount of data collected in a systematic manner, whether of flight parameters, air traffic control, weather, ground and airborne systems, has dramatically increased over the past years. Significant efforts are focusing on the analysis of quantitative data to enhance safety. However, event reports or other text contents that are rich in meaning remain analysed at a local scale, manually, most of the time by experts having in mind a historical overview of events and clustering them on this basis. Yet, contents in natural language embed extremely valuable information to understand how safety of air operations works or fails. Indeed, they reflect especially a variety of elements of the operational context as well as insights of the dynamic of events that support a global understanding of what contributes to or undermines safety.

Objectives

The objective of this thesis is to develop an approach and tools allowing for processing huge sets of text data in natural language(s), that is non-structured, on air transport actual operations in order to derive safety insights in three complementary areas: the understanding of know risks, and identification of mitigation ways forward, the identification and recognition of weak signals and the early detection of emerging risks.

The main scientific challenges are:

- Transforming texts written in natural language, combining several languages, using specific aeronautical vocabulary, into mathematical objects embedding safety relevant elements. Applying standard filters (for example those available in R’s tm package) is not sufficient in this framework. Adapting statistical models called “topic models” to come up with topics that make sense from a safety viewpoint. Based on the corpus, these models generate topics that are distributions of probabilities on the terms of the corpus, and associate with each document/report a distribution of probabilities on the topics. The number of topics is to be optimized to ensure that the most representative terms of each topic can be associated with a relevant interpretation by a safety expert.
- Defining a distance between reports (transformed as aforementioned) allowing for performing a statistical analysis and an automatic classification of event reports leading to clusters (known risks and

weak signals), to identifying outliers and/or emerging clusters (emerging risks). The Wasserstein distance is currently widely used in statistics to perform efficient data analysis for example PCAs and even text analysis.

- Applying methods and algorithms on huge sets of data and presenting the results in a way that is understandable and usable.
- Aggregation of structured data (as trajectories, raw flight data, …) and unstructured data (as texts,…), always in the overall context of Safety, is the ultimate goal.

Skills

- Strong knowledge and interest in Probability and Statistics:
- Strong interest in Air Transport
- Advanced English level
- Programming knowledge (e.g. R, Python will be used in this thesis)
- Open minded is required to interact and work with people with various profiles and background

Support on computational linguistic and aviation safety will be provided by Safety Data and ENAC.

References

J. Bigot, R. Gouet, T. Klein, A. Lopez, *Geodesic PCA in the Wasserstein space*. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques Volume 53, Number 1 (February 2017), 1-26.

D.M. Blei, *Probabilistic topic models.*Communications of theACM, vol. 55(4), pp.77-84, 2012.

A. Genevay, M. Cuturi, G. Peyre, F. BachStochastic *Optimization for Large-scale Optimal Transport* (https://hal.archives-ouvertes.fr/hal-01321664v2).

K. D. Kuhn, *Topics and Trends in Incidents Reports using Structural Topic Modeling to Explore Aviation Safety Reporting System Data*. Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017)

M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, *From word embeddings to document distances*. In ICML,2015.

J. Pennington, R. Socher, and C.D. Manning, *Glove: Global vectors for word representation*. Proc. of the Empirical Methods in Natural Language Processing (EMNLP 2014), 12:1532–1543, 2014.

L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, C. Raynal, *Natural language processing for aviation safety reports: from classification to interactive analysis*. Computers in Industry, vol. 78, pp. 80-95, 2016.

Compétences nécessaires

- Probabilités et statistiques ;
- programmation (R, Python…) ;

Qualification ou formation (Niveau, Diplôme, Certificats, etc.)

- **Bac+5**
- **Profil recherché :** Débutant
- **Déplacements prévus :** Non
- **Conditions d'exercice :** Non

Contrat

- **Date de recrutement prévue :** 1er janvier 2019
- **Durée du CDD :** 3 ans
- **CDD renouvelable**
- **Lieu de travail :** ENAC Toulouse
- **Salaire** (Montant brut mensuel) **:** 2022,98 euros

Coordonnées de la personne chargée de la réception des candidatures

Nom : Bieder Prénom : Corinne

Adresse mail : corinne.bieder@enac.fr

N° de téléphone : 05 62 25 96 25