# Clustering Market Baskets with Bagging and Latent Dirichlet Allocation at Customer and Transactional Levels

Mariflor Vega[1], Ioanna Manolopoulou[1], Ed Manley[1] and Dani Theodoulou[2]
[1]University College London, [2]Sainsbury's

## Project Background

This dissertation investigates the application of Latent Dirichlet Allocation (LDA) in order to cluster market baskets at customer and transactional levels; and introduces a bagging (or bootstrap aggregation) method to improve the stability of topic modelling using data from Sainsbury's. The primary aim was to develop a means to understand the different types of customers based purely on the content of their baskets. Analysing customer behaviours by aggregating their transactions is only possible when customers swipe the loyalty card in exchange for loyalty points for the value of their purchases. However, 57% of transactions are recorded without a loyalty card, preventing the company from having a complete understanding of their customers and their different behaviours. Thus, the complementary need of building comparisons between loyalty and non-loyalty transactions arises in order to determine whether both types of transactions exhibit the same type of behaviours.

## Data and Methods

Topic Models such as LDA were developed in order to uncover the hidden topical patterns in a collection of documents. The documents are defined as bags of words where the grammar and word order are disregarded, and word frequencies are document features. Implementing LDA for retail data not only allows us to discover interpretable topics that characterise different types of market baskets, but also handles the high variety of items. In our interpretation of topic modelling, transactions take the place of documents and the items replace the words, where the order of items do not play a significant role. However, topic model inference inherently produces different realizations of the underlying topic distributions, deeming a global interpretation challenging. We introduce a novel methodology which utilises Bagging in order to improve stability by identifying the topics that appear frequently throughout multiple realizations of LDA.

We implemented LDA and Bagging algorithms on four experiments. First, we identify types of customers through aggregated loyalty transactions. Second and third, we cluster loyalty and non-loyalty transactions independently. Fourth, we cluster both type of transactions in a balanced sample. Subsequently, we analysed the topics across the four experiments in order to identify the type of topics that only characterise either loyalty and non-loyalty transactions and their connection at the customer level.

## Key Findings

We found a variety of topics that describe the type of customers and transactions, from baskets that contain fruits and vegetables to baskets that contain confectionery and snacks. The majority of shopping behaviours exist in both loyalty and non-loyalty transactions. However, there is a set of behaviours that reflect almost exclusively non-loyalty behaviours (these transactions include high proportions of tobacco sales). On the other hand, we have not found super categories that exclusively characterise loyalty behaviours. The implementation of Bagging alongside LDA retrieves seeds that generates topics that appear more frequently throughout multiple versions of LDA. Therefore, this implementation retrieves similar topic distributions for different runs, as opposed to different topic distributions for different realisations. We achieve this by calculating the similarity between analogous topics with and without Bagging over loyalty and non-loyalty transactions. We observed that for both types of transactions, LDA with Bagging retrieved 14% and 35% closer topics, concluding that Bagging improves the stability of LDA.

## Value of the Research

This research developed a practical application of topic modelling in order to cluster market baskets that describe customer behaviours and type of transactions. Some commercial applications of this research might be the development of directed marketing campaigns and a recommender system. The identification of topics that are almost exclusive of non-loyalty transactions could help the retailer tailor their stock to meet the needs of non-card holders. Furthermore, the research contributes to science by introducing a new method that improves consistency and stability on Topic Modelling results.