**Dissertation title**

The use of natural language processing in the topic identification and classification of planning casework records in support of the monitoring of the National Planning Framework 4 in Scotland.

**Introduction**

There is a growing aspiration for the use of machine learning and spatial analysis techniques in the urban planning system but this is not yet deployed on a systematic basis outwith niche areas of innovation. The manageable administrative and spatial scale of the planning system in Scotland offers an opportunity to do so, with the recent publication of the National Planning Framework 4 (NPF4) and an increasing centralisation of locally-applied policies at the regional and national levels, alongside a growth in the availability of development planning and development management data.

One barrier to data use is the inconsistent application of data standards across the 34 planning authorities in Scotland. This has many root causes but these are largely due to local interpretations of national requirements including local procurement of data management systems, with a relatively recent convergence of data and digital platforms.

This is a common issue in casework management and solutions have been developed based upon natural language processing (NLP). An exploration of recently-available planning casework data with a view towards common classification will inform the development of similar solutions in the planning domain.

**Literature review**

The literature in this area is disparate, although the research and use of machine learning (ML) and artificial intelligence (AI) in urban planning is growing and converging, including NLP. From a technical perspective, Tekouabou et al (2021) have considered the application of machine learning methods in urban planning, with a useful taxonomy of methods, models and indicators. However, this does not cover NLP in detail, although it is mentioned in the context of neural networks. Likewise, Chaturvedi and de Vries (2021) provide a good overview of the technical applications, but largely in the context of earth observational data, albeit with a good breakdown of land use indicators, data sources, measurements and applications. Nevertheless, there is a comprehensive overview of ML and AI methods and applications that provides a useful baseline for methodological and model development and a recent overview of the use of machine learning for spatial analyses is also offered by Casali, Aydin and Comes (2022), drawing upon Google Scholar and the Web of Science with a focus upon 2021, identifying the most prominent topics and uses and also knowledge gaps in the sector.

NLP needs to be considered as a detailed subset of ML and AI for a more granular trawl of the literature. A systematic review of NLP for urban research is also provided by Cai (2021) highlighting a limited use of NLP in urban studies, albeit growing in use in recent years, with a comprehensive survey of literature and geographical areas of study. Recent research by Mieczko and Desmond (2023) considers the use of NLP in a national zoning and land use database from publicly available administrative data in the United States. Fu, Li and Zhai (2023) have also looked at NLP in the study of resilience plans from the 100 Resilience Cities network, with useful coverage of topic modelling. Brinkley and Stahmer (2021) have looked in more geographical detail at city planning in California. The most relevant research to this project is by Yai and Konstokosta (2019) which considers the use of Latent Dirichlet Allocation (LDA) for topic modelling of building permit data for seven US cities, using 2.5 million records.

Research in the UK is more limited but there are examples from the UK government. Dray (2019) has considered the use of NLP in government and there is a report by the Turing Institute (2021) researching the use of AI and ML in automating planning applications, which does not consider NLP in detail other than in data cleaning, but does look at the application of AI and ML techniques in the context of the English planning development management system, which is essentially the same as in Scotland.

The above research is largely exploratory, demonstrative or occasional, rather than operational. A systemic example of NLP in a production environment is provided by work by NatureScot (2021) reported on the UK Authority website (2022) that illustrate the progress of the InformedDecision platform. This provides for a further case study in terms of potential capability.

From the perspective of planning in Scotland, there is no literature pertaining to the use of AI and ML in the planning system as it stands. Information is available on planning performance statistics, collected regularly and published quarterly and annually by the Scottish Government. The most recent statistics (2021/22) provide a categorisation of application and development types which, aligned with the topic modelling and UK-specific research mentioned earlier, provides the elements of a taxonomy for NLP and categorisation. However, for minerals, this is generic and does not consider the types of extraction, ie for aggregates.

The recently-published NPF4 (2023) is the overarching land use policy document for Scotland.

**Methodology**

This research employs Python and various libraries throughout the process. This is summarised below in **Figure 1**.

**Figure 1. Methodology**

A screenshot of a computer

Description automatically generated

There are three main stages:

**1. Data Identification, Gathering and Analysis**

This analysis focusses upon planning casework data from 2009 to 2023, for 31 of the 33 planning authorities in Scotland, sourced from the Improvement Service. The years 2020 and 2021 have been omitted due to the Covid-19 pandemic and the stalling of some development activity.

This is a point and polygon dataset of approximately 500,000 records dating back to 2009. This provides, at various scales, both an overview and a detailed record of historical land use development activities. There is an element of language consistency within this dataset. The fields include the general description of development, some high level categorisation by land use and application type, planning decision (approve or refuse) and casework dates.

This stage includes exploratory data analysis, to allow more precise definition of the analytical process, and preliminary visualisations. However, a common taxonomy, vocabulary and categorisation system will be required, which is considered in **Stage 2** below.

**2. Text Mining and Topic Identification**

Following some data cleaning, the most useful fields are subjected to topic modelling using the LDA libraries and methodology. Lai and Kontokosta (2019) provide a good summary of the LDA process, as the creation of summarized N-dimension vector topics based on the probability distribution of words in the corpus, the latency arising from the unknown number of topics.

**3. Data Modelling, Analysis and Visualisation**

The pyLDAvis package allows the visualisation of the LDA outputs and consideration of the definition of topics. This is informed by the lists of categories already employed in development management (E-planning and planning performance framework reporting)

Analysis is conducted in a Python 3.7 environment using Jupyter Notebook and the following libraries:

The notebooks are available at:

**Data and Exploratory Analysis**

Planning casework data (n=588,491) has recently been collected and conflated by the Improvement Service, covering the period 2009 to 2023, for 31 of the 33 planning authorities in Scotland. This has been published as a public sector data service, with an open data version[[1]](#footnote-1), as point and polygon datasets via the Spatial Hub in July 2023.

Each record represents a planning application that has been registered on the Idox Uniform data management system by the appropriate planning authority. This tracks the planning application from validation to conclusion, typically the approval or refusal of the appropriate planning consent.

The data schema used in this analysis is shown in **Figure 2**, and **Figure 3** and **Figure 4** summarise the breakdown of records by year and by planning authority.

**Figure 2. Data Schema**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| year | The calendar year pertaining to the application |
| local\_auth | The planning authority responsible for the application |
| appl\_desc | A local codified description of the application type |
| dev\_desc | A local codified description of the development type |
| stat\_desc | A local codified description of the application status |
| proposal | A free text description of the development proposal |

**Figure 3. Records by year**

|  |  |
| --- | --- |
| **Year** | **Records** |
| 2009 | 31813 |
| 2010 | 43473 |
| 2011 | 42676 |
| 2012 | 40722 |
| 2013 | 41401 |
| 2014 | 42843 |
| 2015 | 43118 |
| 2016 | 42672 |
| 2017 | 42428 |
| 2018 | 41177 |
| 2019 | 40561 |
| 2020 | 37731 |
| 2021 | 42737 |
| 2022 | 39242 |
| 2023 | 16347 |

**Figure 4. Planning Authorities**

|  |  |
| --- | --- |
| **Authority** | **Records** |
| Aberdeen City | 22895 |
| Aberdeenshire | 53130 |
| Angus | 13923 |
| Argyll and Bute | 25664 |
| Cairngorms National Park | 5881 |
| City of Edinburgh | 73221 |
| Clackmannanshire | 3829 |
| Dumfries and Galloway | 23911 |
| Dundee City | 11689 |
| East Ayrshire | 13451 |
| East Dunbartonshire | 12416 |
| East Renfrewshire | 9527 |
| Falkirk | 10464 |
| Fife | 41350 |
| Glasgow City | 36069 |
| Highland | 53734 |
| Inverclyde | 6036 |
| Loch Lomond and Trossachs National Park | 4917 |
| Midlothian | 9631 |
| Moray | 12556 |
| Na h-Eileanan an Iar | 6711 |
| North Ayrshire | 10049 |
| North Lanarkshire | 17378 |
| Orkney Islands | 7370 |
| Perth and Kinross | 27768 |
| Renfrewshire | 11705 |
| Shetland Islands | 6600 |
| South Ayrshire | 15778 |
| South Lanarkshire | 25697 |
| Stirling | 11764 |
| West Dunbartonshire | 3828 |

Although there are codified descriptions for **appl\_desc**, **dev\_desc** and **stat\_desc**, these are subject to considerable variation both across the range of planning authorities and also across the time period of the dataset. This is shown in **Figure 5** below and **Figure 6** provides a summary of the types of development defined in Planning Performance Framework statistical returns submitted annually to the Scottish Government and also the categorisation employed by the national E-Planning portal, launched in 2016.

**Figure 4.**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Number of discrete descriptions** | **Predominant description** |
| appl\_desc | 643 | “Full Planning Permission”: 103,467 instances |
| dev\_desc | 681 | “Householder Developments”: 68,572 instances |
| stat\_desc | 466 | “Application Approved”: 89,420 instances |
| proposal | 388,788 | ALTERATIONS AND EXTENSION TO DWELLINGHOUSE”: 76 instances |

**Figure 5. Planning performance Framework, Use Classes and E-Planning Categories**

|  |  |  |  |
| --- | --- | --- | --- |
| **PPF Local/Major** | **PPF Category (dev\_desc)** | **Use Classes (proposal)** | **E-Planning Category (app\_desc)** |
| Local | Approval of Matters Specified in Conditions | Class 1. Shops | Advertisement Consent |
| Local/Major | Business & Industry | Class 2. Financial, professional and other services | Planning permission |
| Local/Major | Electricity Generation | Class 3. Food and drink | Listed Building Consent |
| Local/Major | Fish Farming | Class 4. Business | Conservation Area Consent |
| Local | Householder | Class 5. General industrial | Householder |
| Local/Major | Housing | Class 6. Storage or distribution | Certificate of Lawfulness (Existing) |
| Local/Major | Minerals | Class 7. Hotels and hostels | Certificate of Lawfulness (Proposed) |
| Local/Major | Other Developments | Class 8. Residential institutions | Prior Notification |
| Local | Telecommunications | Class 9. Houses | Tree Works |
| Local/Major | Waste Management | Class 10. Non-residential institutions |  |
|  |  | Class 11. Assembly and leisure |  |

It is clear, from Figure 5 above, that there are ten PPF categories that could potentially be applied to the **dev\_desc** field, eleven classes from the use classes order that might apply to **app\_desc** and/or **proposal**, and nine E-planning categories applicable to the **app\_desc** field. Applied to the 31 authorities in this instance, that would reduce the total of **dev\_desc** categories from 681 to 10 and the app\_desc categories from 634 to 9. Even if each authority categorisation (from the above) is considered as unique across Scotland (ie. “Aberdeen City – Householder”, there would still be a reduction to 310 and 279 categories respectively, a reduction of over 50%.

This is tested using the LDA process. However, the list above is unlikely to be granular enough for the expected purposes of a digital planning system, notably the wide-ranging category for “business and industry”. Also, there are no quantities and units of development. Nevertheless, there is a significant amount of detail locked up in the **proposal** field.

**The *Proposal* field**

The **proposal** field is a free-text field that provides a concise yet detailed description of the type of development. This is defined by an applicant but modified by the planning case officer, leading to an inherent consistency in the use of the language. (The scope for large language model automation of this process is interesting but beyond the scope of this analysis). Some examples of this field are shown at **Figure 6** below.

**Figure 6. Examples of the proposal field**

|  |
| --- |
| Change of use from Light Industrial and Storage and Distribution (Class 4 and Class 6) to the sale of motor vehicles - In restrospect. |
| Change of use from domestic accommodation ancillary to 26 Milton Road East to a separate dwelling, form two dormers and new window openings |
| Major' redevelopment of the Ellengowan Drive estate, replacing existing 128 cottage flats with a mix of houses, flats and cottage flats for mid-market and social rent |
| Erection of 132 dwellings with associated roads, SUDS, landscaping & ancillary works, formation of vehicular accesses to the A8 Glasgow Road and Hillwood Rise (Amended description) |

The semi-standard nature of the description is evident although still with scope for confusion and clutter. There is also some overlap with the **app\_desc** and **dev\_desc** fields. Nevertheless there is potential to enrich the dataset by extracting useful information as data, which is considered as part of the subsequent data process, focussing on quantities and units of development and also defined use classes.

**Data Preparation and LDA Modelling**

The principal method of analysis is Python (3.7) in Jupyter Notebook. The following steps are employed:

1. Loading of a shapefile as **planapp\_gdf** geodataframewith the removal of 2020 and 2021 as Covid years and also the erroneous year “3016”.

2. Selection of a random subset of data (***n*** = )records within computational processing limitations.

3. Definition of columns subject to analysis: **dev\_desc**, **app\_desc**, **proposal**.

4. Conversion to all strings to lower case.

5. Extraction of use classes, quantities and units of development from the **proposal** field for subsequent use.

6. Conversion of text numbers to numeric and removal of numeric values.

7. Data cleaning and text pre-processing including the identification and removal of stopwords (an extended list including some common planning terms), words of string length less than four (with some exceptions) and non-alpha characters, via regular expressions, followed by stemming and lemmatization of verbs and nouns.

8. Visualisation of word counts as bar charts and word clouds.

9. Creation of corpus and LDA modelling (coherence using *u\_mass* for computational reasons) using defined number of topics for **dev\_desc**, **app\_desc** and **proposal**.

10. Visualisation of coherence to confirm the appropriate number of topics in each case, informed by the earlier **Figure 5**.

11. Topic modelling and visualisation via pyLDAvis.

12. If time allows – classification modelling and testing using Random Forest or similar

13. Analysis and conclusions.

11.

1. <https://data.spatialhub.scot/dataset/planning_applications_ccf-is> [↑](#footnote-ref-1)