Demonstrator: Text-based Access to Olfactory Information

Deliverable D3.6

Version FINAL

NEGOTIATING OLFACTORY AND SENSORY EXPERIENCES IN CULTURAL HERITAGE PRACTICE AND RESEARCH

The Odeuropa project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004469. This document has been produced by the Odeuropa project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Grant Agreement No.	101004469
Project Acronym	ODEUROPA
Project full title	Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research
Funding Scheme	H2020-SC6-TRANSFORMATIONS-2020
Project website	http://odeuropa.eu/
Project Coordinator	Prof. Dr. Inger Leemans KNAW Humanities Cluster Email: inger.leemans@huc.knaw.nl
Document Number	Deliverable D3.6
Status & version	FINAL
Contractual date of delivery	31 December 2023
Date of delivery	30 December 2023
Type	Demonstrator
Security (distribution level)	Public
Number of pages	16
WP contributing to the deliberable	WP3
WP responsible	Sara Tonelli
EC Project Officer	Hinano Spreafico
Authors: Stefano Menini[1], Elisa Leonardelli[1], Sara Tonelli[1]	
Internal reviewers: Mathias Zinnen,[2], Ali Hürriyetoğlu[3], Inger Leemans [3]	
Affiliations (1) FBK, (2) FAU, (3) KNAW	
Keywords: Olfactory information extraction, Python notebook	
Abstract: The current document contains a brief description of the functionalities of the demonstrators for text-based access to olfactory information, which represent the actual deliverable 3.6. The first demonstrator, the *Odeuropa Text Nosebook – Textual Smell Trend Analyser*, is a Python notebook which allows users to interactively analyse olfactory information in multiple languages in context and across time, including their relations to emotions, starting from the output of previously processed corpora. The second demonstrator, the *Smell Reader*, can be used through a user-friendly interface and it gives access to the underlying system for olfactory information extraction in six languages which can be used to process the users' text snippets in real time, exporting the output analysis also in Excel format.	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	11 December 2023	Template setup and first draft	Sara Tonelli	All
0.2	14 December 2023	Document ready for revision	Sara Tonelli	All
0.3	15 December 2023	First review round	Mathias Zinnen	All
0.4	15 December 2023	Second review round	Ali Hürriyetoğlu	All
0.5	18 December 2023	Final corrections	Sara Tonelli	All
1.0	30 December 2023	approval by project manager	Marieke van Erp	-

Executive Summary

Summary table	
Challenges	The challenges associated with this kind of demonstrators is addressing the needs of users with some technical skills, that may be interested in using Odeuropa system for research purposes, as well as of users with no specific research interests and no knowledge of Python. Such different requirements have been addressed through the development of two different demonstrators: one, which we call the *Odeuropa Nosebook*, is based on a Python notebook, while the other, the *Smell Reader*, displays a simple, more intuitive GUI.
Barriers	One general barrier faced when building the demonstrators is the need to develop tools that, despite processing large amounts of data, would ideally display the results in a short time. This was addressed by giving the possibility to process one language at a time in the Python notebook, therefore downloading the different resources separately. As regards the *Smell Reader*, a limit was set to the length of the documents that can be processed on the fly.
Practices	Odeuropa demonstrators follow best practices in terms of privacy. In particular, to access the *Odeuropa Nosebook* a Google account is required, so that the processing is run on the user account, where the output is saved. Each user is free to modify the notebook, add new analyses, etc.
Guidelines	While the *Smell Reader* does not require specific guidelines, the *Odeuropa Text Nosebook* has been enriched with detailed explanations on how to launch the analyses and tailor them to users' needs.

Layman's Summary

This document supports the release of deliverable D3.6, which encompasses two different demonstrators. The first one is the *Odeuropa Nosebook – Textual Smell Trend Analyser*, which includes interactive notebooks in Python for analysing multilingual olfactory information from text in context, across time and in relation to emotions. The second one is the *Smell Reader*, which enables real-time processing of textual data in six languages to extract olfactory information. Both demonstrators can be accessed online and the output of the processing can be exported by users for further analysis. Furthermore, both support six languages: Dutch, English, French, German, Italian and Slovene.

Contents

1 Introduction

The last deliverable of WP3 related to text-based analysis of olfactory information is devoted to making the software outputs as much accessible as possible, also to support impact and dissemination activities beyond the project end. This complements D2.6. [Zinnen and Christlein, 2023], which pursues the same goals related to image analysis. We present two different demonstrators, which do not require programming skills to be run, although the first one addresses specific research questions in the digital humanities / computational linguistics community, while the second one is designed mainly for laypeople:

- *The Odeuropa Nosebook – Textual Smell Trend Analyser* which enables users to analyse smell sources, qualities and related emotions extracted from the Odeuropa historical corpora in six languages, and
- *the Odeuropa Smell Reader*, which lets users paste arbitrary text snippets to try out olfactory information extraction on their own.

Both tools will be introduced and further discussed in the following sections.

2 Odeuropa Nosebook – Textual Smell Trend Analyser

To access the Odeuropa Nosebook it is necessary to open the Jupyter notebook available at: https://bit.ly/odeuropa-text. Since the notebook is based on Google Colab, users can perform data analysis through the browser, without the need to install locally any Python library. However, a Google gmail account is required.

The notebook is introduced by the short overview displayed in Figure 1, where the different sections are summarised. They will be explained in more detail in the following subsections. Before performing any analysis, however, the "Load Functions" cell has to be run by clicking on *12 cells hidden* to set the environment ready for further processing. Then, the language of interest has to be selected by clicking on the arrow before *language_selection()*. This step will retrieve the data in a specific language that will be analysed in the following steps.

The downloaded data contain, for each language, the smell word, smell source and quality that have been automatically identified using the Odeuropa system for olfactory information extraction (v2) described in D3.5 [Novalija et al., 2023]. Indeed, the notebook is not meant to extract olfactory information from new data, but to explore what has been previously extracted from the Odeuropa historical corpora (see https://github.com/Odeuropa/benchmarks_and_corpora). Note that downloading the data may take some minutes, especially for English, Slovene Dutch and Italian, because also the emotion detection model has to be downloaded.

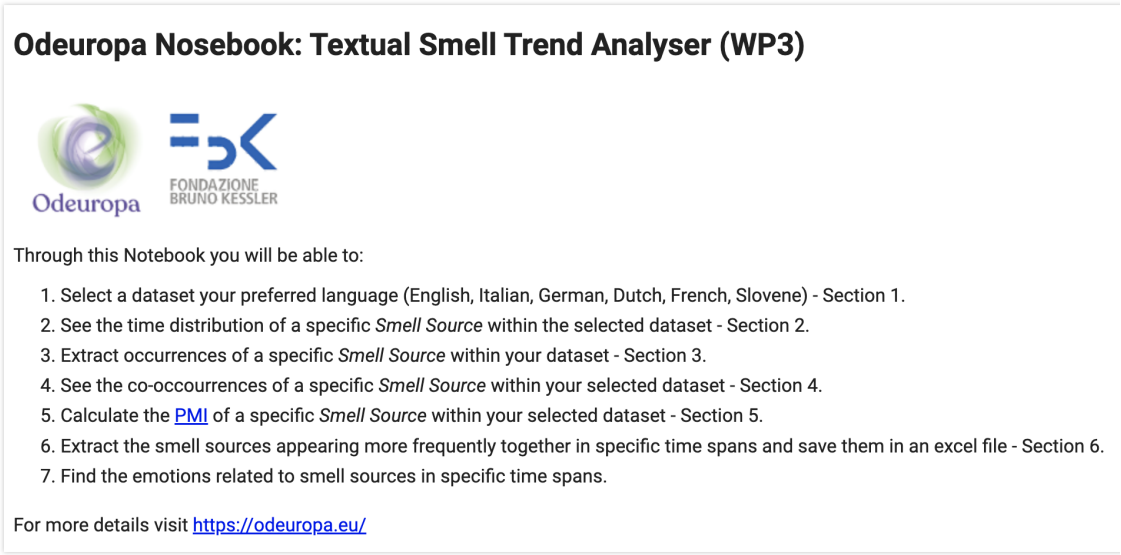

Figure 1: Overview of Jupyter notebook sections

2.1 Smell Source Distribution over time

The first analysis that can be carried out on the data after setting up the notebook and selecting the language of interest is to display the distribution over time of a given smell source. We report in Figure 2 a screenshot of the notebook command to perform this analysis. Running the *run_extract_examples()* cell displays the graphical interface, where it is possible to type a smell source, define one or more time spans of interest and then display a graph with the number of occurrences over time.

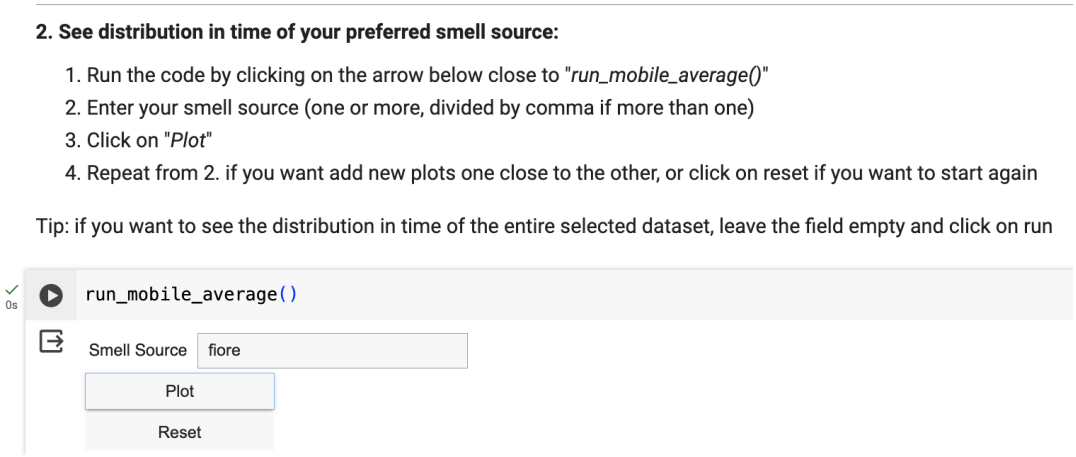

Figure 2: Distribution over time of a smell source

As an example, we report in Figure 3 the occurrences of the term “fiore” (flower), which has been queried in the Italian data. Note that the number of occurrences is not normalised, therefore the output is influenced by the number of documents available for each time period.

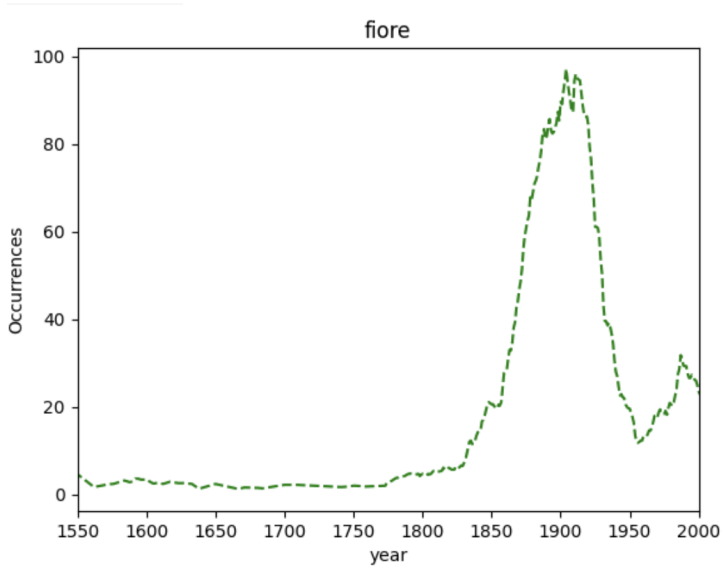

Figure 3: Occurrences of the smell source “fiore” (flower) over time

2.2 Smell Source textual context in different time spans

A second type of analysis that can be performed through the text-based Odeuropa Nosebook is comparing what has been extracted by the olfactory information system related to one or more smell sources. In particular, a user can define a smell source and one or more time spans (see Figure 4) and click on *Show Selection*. In this way, we can extract the list of sentences retrieved from the Odeuropa corpus in which the search term appears as smell source, together with the year of publication of the source document, the name of the document collection (e.g. Wikisource) and the related quality, if present. After the extraction is completed, the output table is displayed and can be exported for further processing by clicking on *Export Excel*.

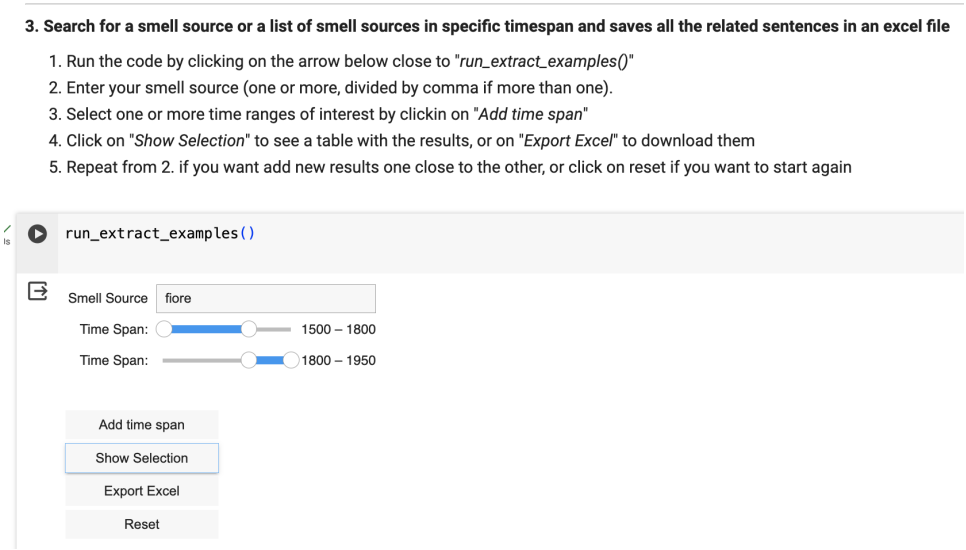

Figure 4: Smell source comparison in different time spans

This analysis can be useful when a close reading of smell sources is required, for instance when it is useful to check the textual context in which they have been mentioned (also to be aware

of possible system mistakes), when it is required to trace a mention back to its original textual collection, etc.

2.3 Word co-occurrences with Smell Sources

This step allows users to extract which terms (excluding stop words) co-occur more frequently with a given smell source. The interface to run the command and define the different parameters is displayed in Figure 5. The number of co-occurring terms is set to 10 but can be increased if needed. Also the window size, i.e. how many words surrounding the given smell source should be considered, can be defined by the user. One or more time spans can also be selected, so to enable comparisons of co-occurrences between different periods.

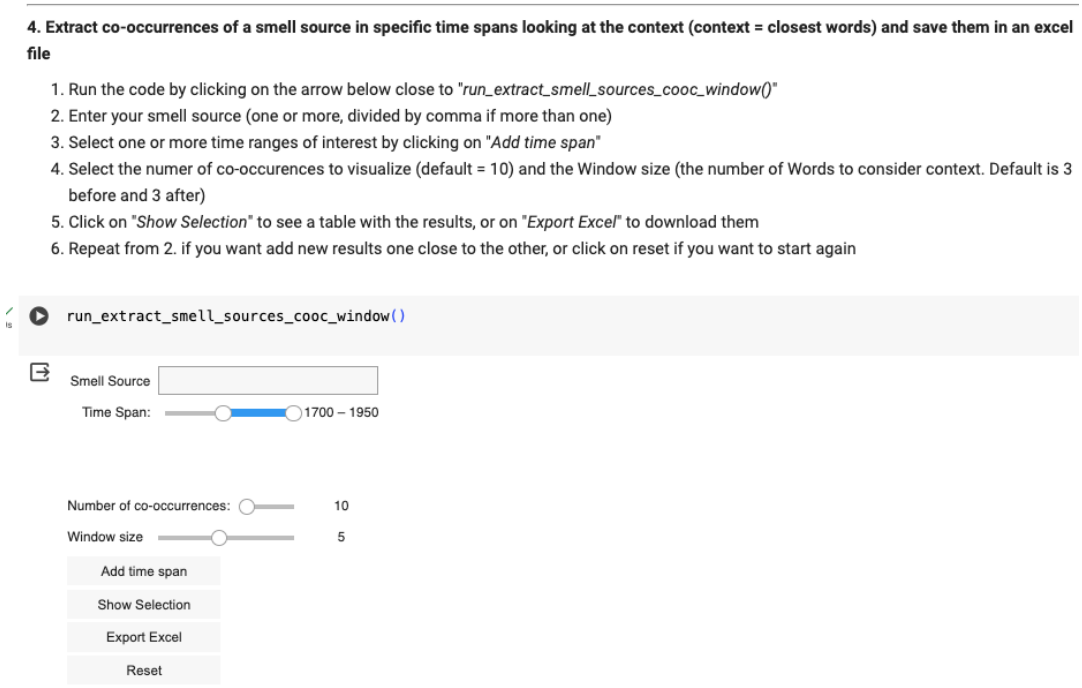

Figure 5: Extraction of terms co-occurring with a given smell source

After setting the above parameters and typing a smell source of interest in the corresponding box, *Show Selection* should be clicked to display one or more lists of terms occurring in the context window of the given smell source. The terms are ranked by normalised frequency, whose value is also reported in a column of the table. As in the previous analysis, also in this case the results can be exported in an Excel file and stored locally.

An example output for the smell source “flower” in the time period 1700–1950 is displayed in Figure 6. We observe that there are no specific requirements concerning grammatical category of the co-occurring terms. In this case, the top-ranked words are verbs and show that flower mentions are present in textual descriptions involving all senses (e.g. “see”, “smell”).

Results for 1700–1950 span, and targets ['']

1 to 10 of 10 entries Filter

Word	Frequency over 1000 words
say	0.36995148238741005
smell	0.1768428729742474
one	0.1076429769104228
come	0.1023190039286102
man	0.09962763568835616
go	0.09333994132793731
see	0.08615904178734682
like	0.07914268004310374
know	0.07693317249215281
well	0.0692821649619983

Show 25 per page

Figure 6: List of terms co-occurring with ”flower” ranked by normalised frequency

2.4 PMI-based extraction of Qualities

Another analysis available in the text-based Odeuropa Nosebook allows users to identify which qualities have been most associated with a smell source, and make comparisons across different time spans (Figure 7). The association strength is computed used pointwise mutual information, a standard measure for studies in semantic shifts. However, different from [Hamilton et al., 2016], we only compute PMI of qualities and not of any term in the surrounding context of the smell source.

The PMI between each smell source (w_1) extracted by our system and its associated qualities (w_2) was therefore pre-computed in the following way:

$$PMI\,(w_1; w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

where $P(w_1, w_2)$ is the probability of the smell source and a word/quality to co-occur, while $P(w_1)$ and $P(w_2)$ are their independent probabilities.

The output is displayed in tabular format: for the given smell source and the selected time period(s), a ranked list of qualities with the corresponding PMI is given, with the possibility to export it in Excel format. This kind of analysis has been used to track the perception shifts of different English smell sources in [Paccosi et al., 2023]

For the “flower” smell source in the time period 1700–1950, for instance, the ranked list of qualities includes “pleasing”, “combustible”, “stimulant” and “succulent”. We observe a deviation from the analysis of raw co-occurrences (Section 2.3) based on the same smell source and time span. This exemplifies how the PMI-based analysis of qualities can provide an additional perspective on historical descriptions of the same smell source.

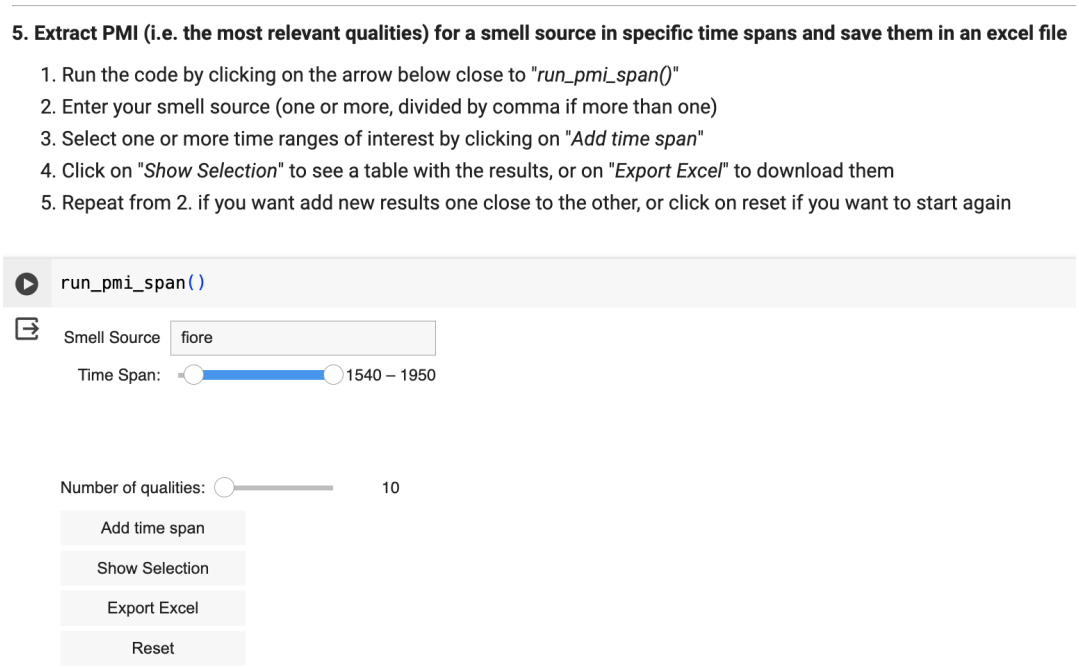

Figure 7: PMI-based extraction of most relevant qualities for a given smell source

2.5 Co-occurrences between Smell Sources

With this function it is possible to investigate which smell sources tend to occur together in different time periods. This information can be used to track a change in the olfactory scene in which a smell source tends to appear, for example when an object first perceived as exotic (therefore appearing with other goods like spices) becomes a commodity (therefore being mentioned with smell sources belonging to everyday life). Again, two parameters can be set: the number of co-occurring smell sources to be included in the output list and the size of the textual window in tokens, in which co-occurrences should be found (see Figure 8).

As in the other analyses, the output is one or more lists containing the smell sources co-occurring with the one of interest, ranked by frequency. If we run this analysis using “flower” and comparing the two time periods 1500–1700 and 1700–1950, we observe that the co-occurring smell sources in the two lists largely overlap, showing that the characterisation of flowers has remained quite stable over time.

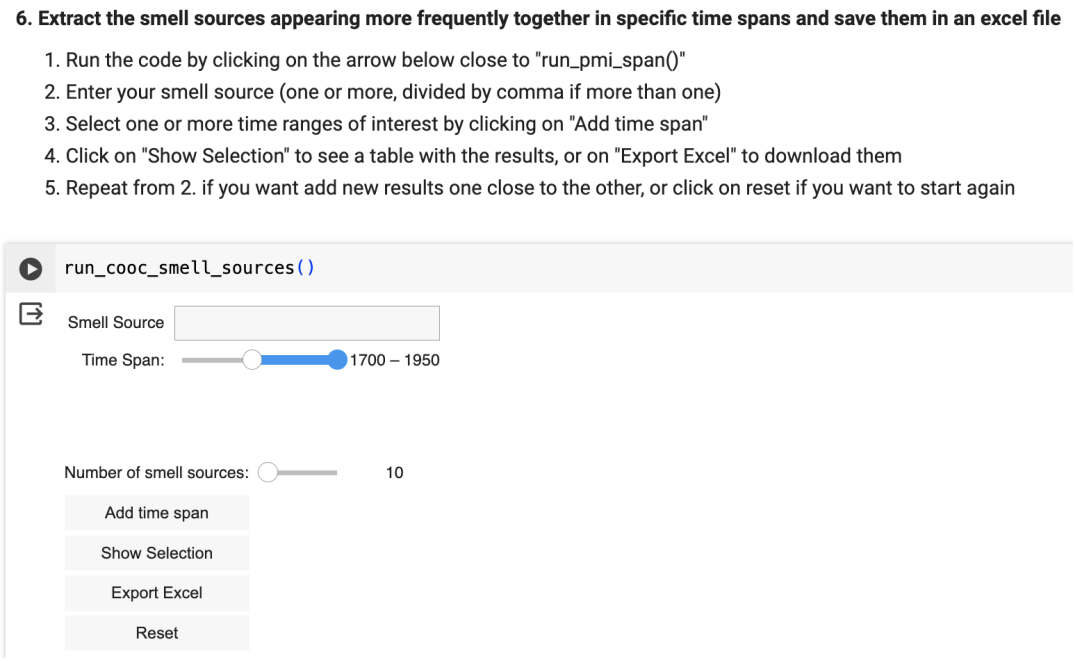

Figure 8: Extraction of smell words most frequently co-occurring with a given smell word

2.6 Emotions related to Smell Sources

The last analysis provided through the text-based Odeuropa Nosebook is the detection of emotions associated with a given smell source. Again, it is possible to define a smell source of interest and set one or more time spans to which the extraction can refer (see Figure 9). The emotion analysis is available in English, Slovene, Dutch and Italian and is performed using the Odeuropa models available at https://huggingface.co/lrei/xlm-roberta-base-emolit-multilingual and https://huggingface.co/lrei/roberta-large-emolit. The details of the implemented approach for emotion detection are reported in D3.5 [Novalija et al., 2023].

After clicking on “Show Emotions”, the emotion analysis model is called and retrieves a list of emotions associated with the given smell source, ranked by relevance in tabular format. For instance, if we run this analysis on “flower”, we observe that the first-ranked emotions associated with this smell source are all positive: approval, gratitude, admiration, relief, amusement and joy.

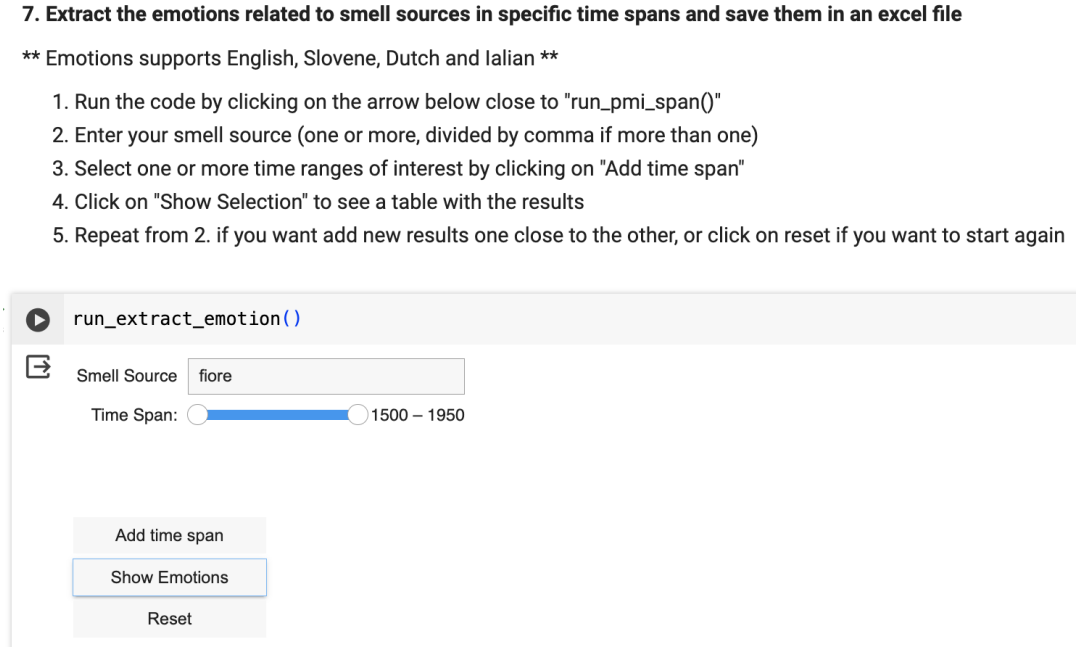

Figure 9: Emotion analysis for a given smell source

3 Odeuropa Smell Reader

The *Odeuropa Smell Reader* is freely accessible at https://smell-extractor.tools.eurecom.fr/ and gives users the possibility to launch olfactory information extraction on texts in six languages: Dutch English, French, German, Italian, and Slovene. Differently from the Odeuropa Nosebook, this demonstrator can be used to process new data on the fly and extract from them olfactory-related information. The underlying system is based on the multi-task learning models for olfactory information extraction presented in D3.5 [Novalija et al., 2023] and available at https://zenodo.org/communities/odeuropa.

By opening the above link, the interface in Figure 10 is displayed.

The user can either insert a text to be processed in the above window (with a limit of 1,000 words), or select one of the examples reported below. Then, the language of the text should be manually set. This step is required to preload the classification model that performs best for each specific language, following the evaluation reported in D3.5 [Novalija et al., 2023]. By clicking on “Extract the smells”, the text is processed on the fly in few seconds by identifying smell words and the related frame elements, if present, based on the annotation guidelines presented in D3.1. [Tonelli and Menini, 2021].

As an example, we display in Figure 11 the system output after selecting the English example available through the interface. The smell words, triggering the olfactory events, are highlighted in light blue, while the other frame elements are marked in different colours: yellow for Locations, red for Perceiver, grey for Circumstances, purple for Quality, orange for Odour Carrier and pink for Effect. By clicking on “Download Output”, an Excel file is stored locally on the user’s computer with one row for each smell event, containing all the annotated elements as well as the current sentence, the one before and the one after.

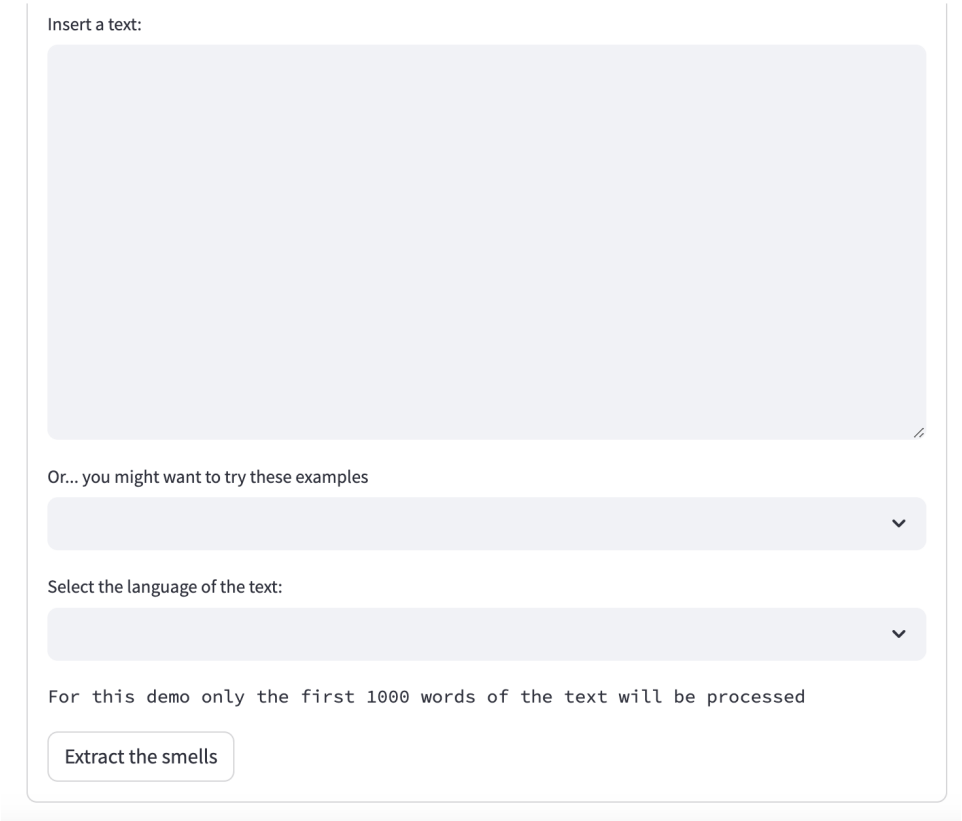

Figure 10: User Interface of *Odeuropa Smell Reader*

4 Conclusions

We present in this document the two demonstrators made available for the extraction and analysis of olfactory information from text: the *Odeuropa Nosebook – Textual Smell Trend Analyser* (Section 1) and the *Odeuropa Smell Reader* (Section 3). We provide all the instructions to access them and explore what has been extracted from the Odeuropa historical corpora, as well as to detect information from new text snippets provided by the user. The release of the two demonstrators is meant to support exploitation activities beyond the end of the project (see D7.10) [Leemans et al., 2023] and can be easily used for disseminating the project results with both researchers and the general public. A first presentation of the tools has already taken place during the Smell Culture Fair as part of the Data Explorations Session (see D7.6) [Leemans and Ehrich, 2023], obtaining a positive feedback from the audience.

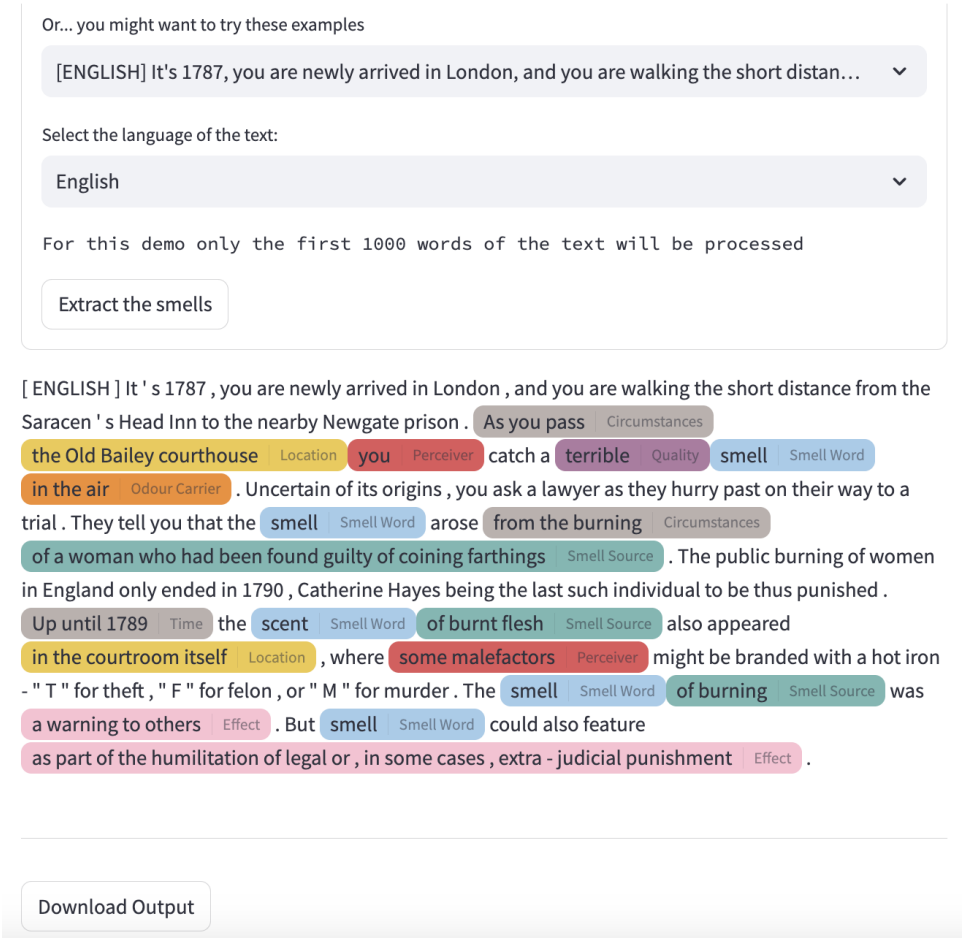

Figure 11: Example analysis on English snippet

References

[Hamilton et al., 2016] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

[Leemans et al., 2023] Leemans, I., Bosse, A., Tullett, W., Christlein, V., and Ehrich, S. (2023). Dissemination and exploitation plan version 2. Technical report, Odeuropa. Deliverable D7.10.

[Leemans and Ehrich, 2023] Leemans, I. and Ehrich, S. (2023). Impact activities report y3. Technical report, Odeuropa. Deliverable D7.6.

[Novalija et al., 2023] Novalija, I., Rei, L., Mladenić, D., Brank, J., Menini, S., Tonelli, S., Hürriyetoğlu, A., and van Erp, M. (2023). Final odeuropa text processing system with emotion analysis version 3. Technical report, Odeuropa. Deliverable D3.5.

[Paccosi et al., 2023] Paccosi, T., Menini, S., Leonardelli, E., Barzon, I., and Tonelli, S. (2023). Scent and sensibility: Perception shifts in the olfactory domain. In Tahmasebi, N., Montariol, S., Dubossarsky, H., Kutuzov, A., Hengchen, S., Alfter, D., Periti, F., and Cassotti, P., editors, *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 143–152, Singapore. Association for Computational Linguistics.

[Tonelli and Menini, 2021] Tonelli, S. and Menini, S. (2021). FrameNet-like annotation of olfactory information in texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–20, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

[Zinnen and Christlein, 2023] Zinnen, M. and Christlein, V. (2023). Demonstrator: Image retrieval and visualization. Technical report, Odeuropa. Deliverable D2.6.