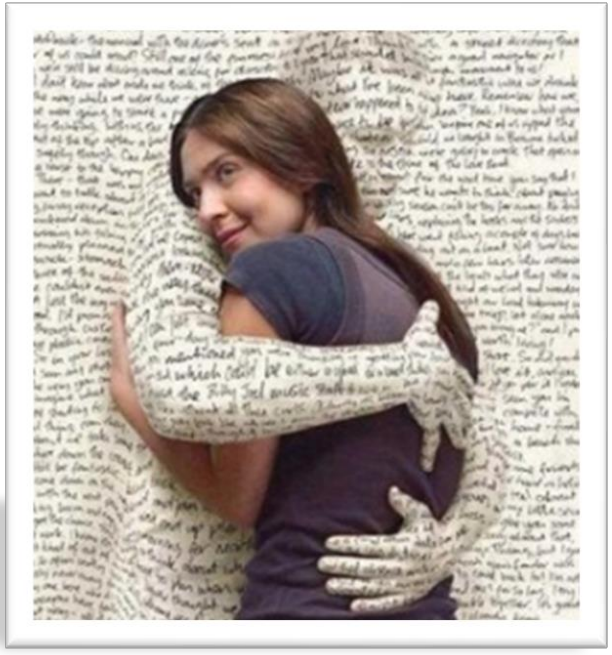
Introduction to NLP

Zurich University of Applied Sciences

zh aw School of Engineering

Lecture 04



Lecturers: Mark Cieliebak and Jan Deriu

NLP Workbench

Anonymous ACL submission

Abstract

This document outlines a comprehensive system of tools for Natural Language Processing (NLP). The toolkit encompasses essential functions and methods for basic text processing, serving as a foundation for more advanced algorithmic implementations. Additionally, it includes tools for analyzing the characteristics of a given text to facilitate comprehension of the corpus and its content.

1 Introduction

Natural Language Processing (NLP) is a prominent field within Data Science that has experienced significant development in recent years, getting considerable interest from both experts and general users. In the pursuit of advancing NLP capabilities, a comprehensive workbench tailored to the needs of practitioners is very important. This one-pager introduces a suggested workbench for NLP implemented using Python.

2 Contribution

The following tools are essential for laying the groundwork in NLP projects, serving as the first steps of text processing before the implementation of more complicated algorithms. Python, as the primary programming language, provides a versatile and intuitive environment for NLP development, while libraries such as NLTK and spaCy offer a rich array of functionalities for text analysis and manipulation.

2.1 Basic Features

In the realm of basic text processing, several fundamental features are indispensable:

Character Count and Visualization: Understanding the length of text, as exemplified in Figure 1, aids in grasping its complexity. Word Count and Diversity. Character Set Analysis, Text Segmentation, Inverted Indexing: Generating an inverted index facilitates efficient word-based search operations. Corpus Statistics: Extracting key statistics, such as file count and vocabulary size. Data Visualization, Corpus-wide Search Capabilities.

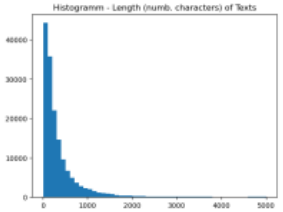


Figure 1: Histogram of lengths of texts

2.2 Further Features

In addition to the basic features, the workbench offers further functionalities to enhance text processing:

Understanding the structure of the text through sentence segmentation. Advanced Search Queries: Enabling precise searches with logical operators. Statistic Filtering, Approximate Search: Implementing fuzzy search capabilities to accommodate for spelling variations. Keyword Extraction: Identifying the most significant words within the text. Topic Modeling: Analyzing and identifying the most relevant topics present within the corpus.

3 Conclusion

The described workbench for NLP presents a comprehensive suite of essential tools tailored for text processing and analysis. By offering a wide array of functionalities, ranging from basic text statistics to advanced search capabilities, the workbench equips users with the necessary resources to efficiently preprocess and explore textual data.

Sample System Description

NLP Workbench

Anonymous ACL submission

Abstract

This document outlines a comprehensive system of tools for Natural Language Processing (NLP). The toolkit encompasses essential functions and methods for basic text processing, serving as a foundation for more advanced algorithmic implementations. Additionally, it includes tools for analyzing the characteristics of a given text to facilitate comprehension of the corpus and its content.

1 Introduction

Natural Language Processing (NLP) is a prominent field within Data Science that has experienced significant development in recent years, getting considerable interest from both experts and general users. In the pursuit of advancing NLP capabilities, a comprehensive workbench tailored to the needs of practitioners is very important. This one-pager introduces a suggested workbench for NLP implemented using Python.

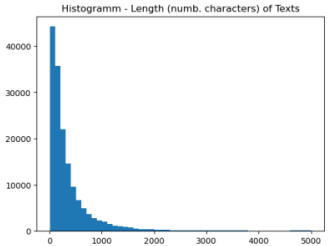


Figure 1: Histogram of lengths of texts

2 Contribution

The following tools are essential for laying the groundwork in NLP projects, serving as the first steps of text processing before the implementation of more complicated algorithms. Python, as the primary programming language, provides a versatile and intuitive environment for NLP development, while libraries such as NLTK and spaCy offer a rich array of functionalities for text analysis and manipulation.

2.1 Basic Features

In the realm of basic text processing, several fundamental features are indispensable:

Character Count and Visualization: Understanding the length of text, as exemplified in Figure 1, aids in grasping its complexity. Word Count and Diversity. Character Set Analysis, Text Segmentation, Inverted Indexing: Generating an inverted index facilitates efficient word-based search operations. Corpus Statistics: Extracting key statistics, such as file count and vocabulary size. Data Visualization, Corpus-wide Search Capabilities.

2.2 Further Features

In addition to the basic features, the workbench offers further functionalities to enhance text processing:

Understanding the structure of the text through sentence segmentation. Advanced Search Queries: Enabling precise searches with logical operators. Statistic Filtering, Approximate Search: Implementing fuzzy search capabilities to accommodate for spelling variations. Keyword Extraction: Identifying the most significant words within the text. Topic Modeling: Analyzing and identifying the most relevant topics present within the corpus.

3 Conclusion

The described workbench for NLP presents a comprehensive suite of essential tools tailored for text processing and analysis. By offering a wide array of functionalities, ranging from basic text statistics to advanced search capabilities, the workbench equips users with the necessary resources to efficiently preprocess and explore textual data.

Reviews

Structure:
Formal requirements: The document meets the one-page requirement, which is commendable. The inclusion of a figure is a positive aspect, but its integration into the text needs to be better handled. There's a lack of explicit reference to Figure 1 within the text, which could confuse readers about its relevance.
Structure critique: The paper could improve its clarity by more explicitly connecting sections and features described to the figure presented. Moreover, while the sections are present, they seem to provide a list of features rather than a coherent narrative about the workbench's development and unique advantages.

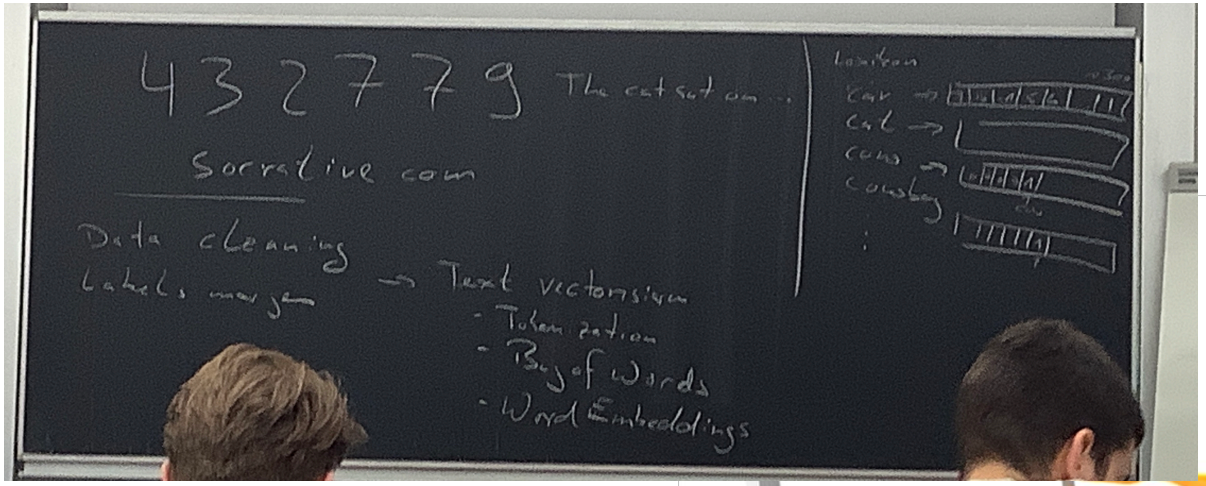
Presentation:
Quality of writing: The prose is academically sound, yet it tends to be dense and may benefit from concise language or examples that could illustrate the workbench's applications more vividly.
Clarity and understandability: The technical elements are adequately described, but the paper lacks a layman's summary that would make the content accessible to a broader audience. There is no discussion or analysis of the histogram in Figure 1, which is a missed opportunity to demonstrate the workbench's practical utility.

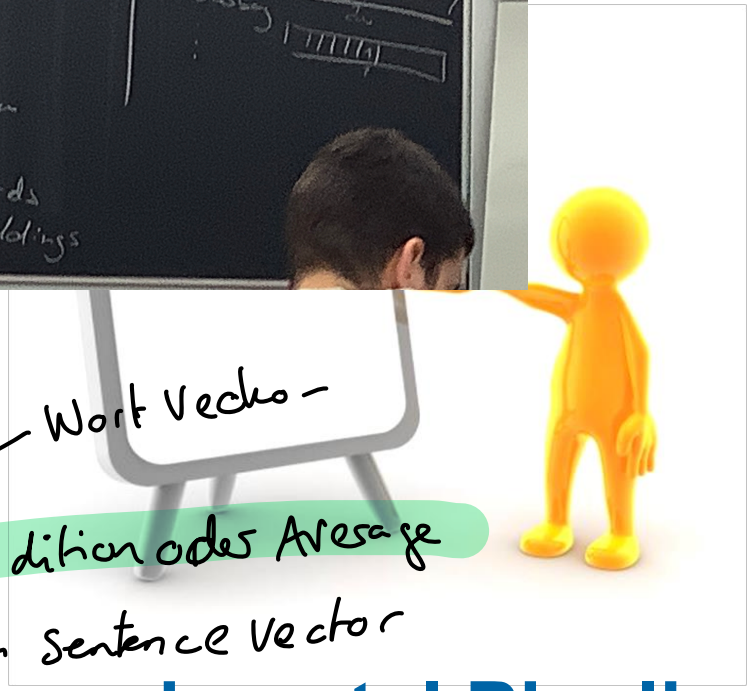
Remarks:
The paper could be enhanced by providing examples of how the workbench has been used in practical scenarios. Details on the ease of use, scalability, and any limitations of the tool would offer a more rounded view of its applicability.

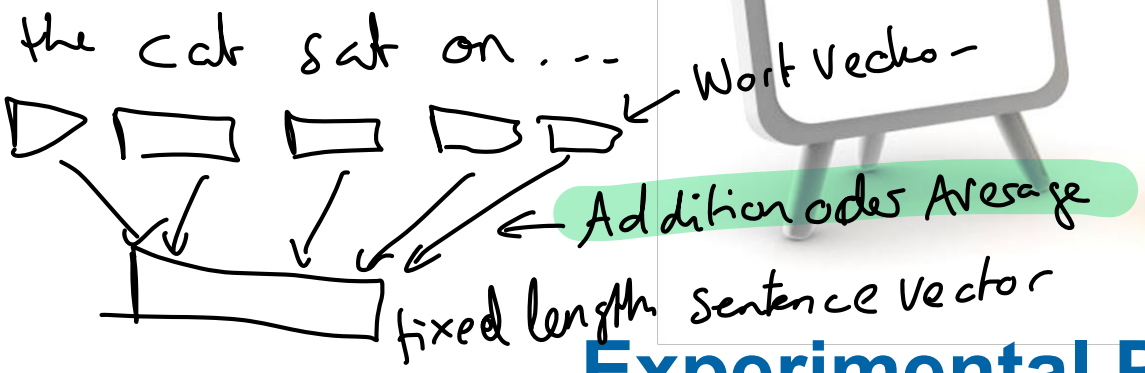
Structure:
Formal requirements are fulfilled
The document is well structured
Figure is labeled and refferenced as well

Presentation:
Clear and coherent, easy to understand.

Remarks:
Including code snippets, implementation details or use cases
Providing benchmarks or comparative analyses







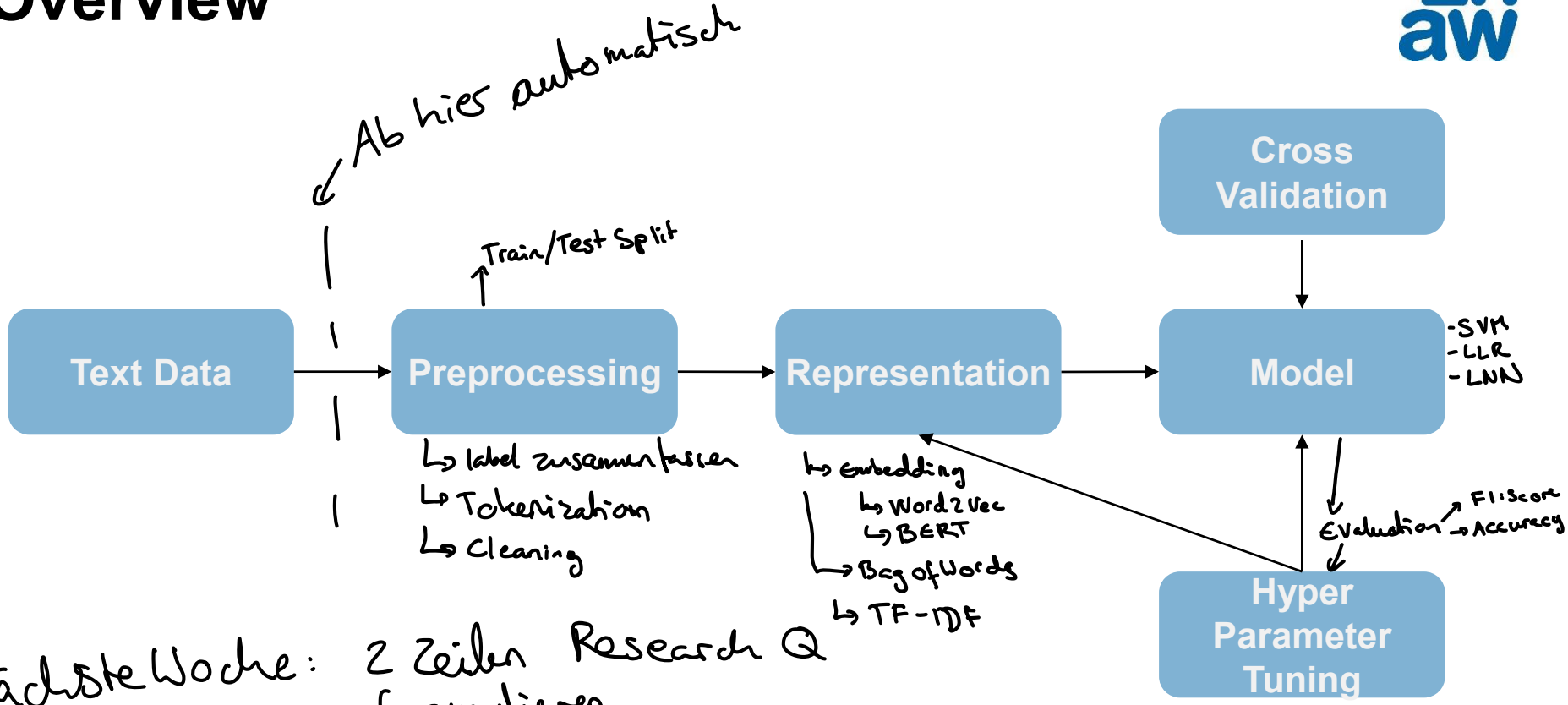
Experimental Pipeline

Data Cleaning
label merging
Text vectorisieren
- Word Embedding
- Bag of words
- Vector of sentence
- Model fill dvector with sentence and label
- train + validation set
→ Validation: F1, Accuracy

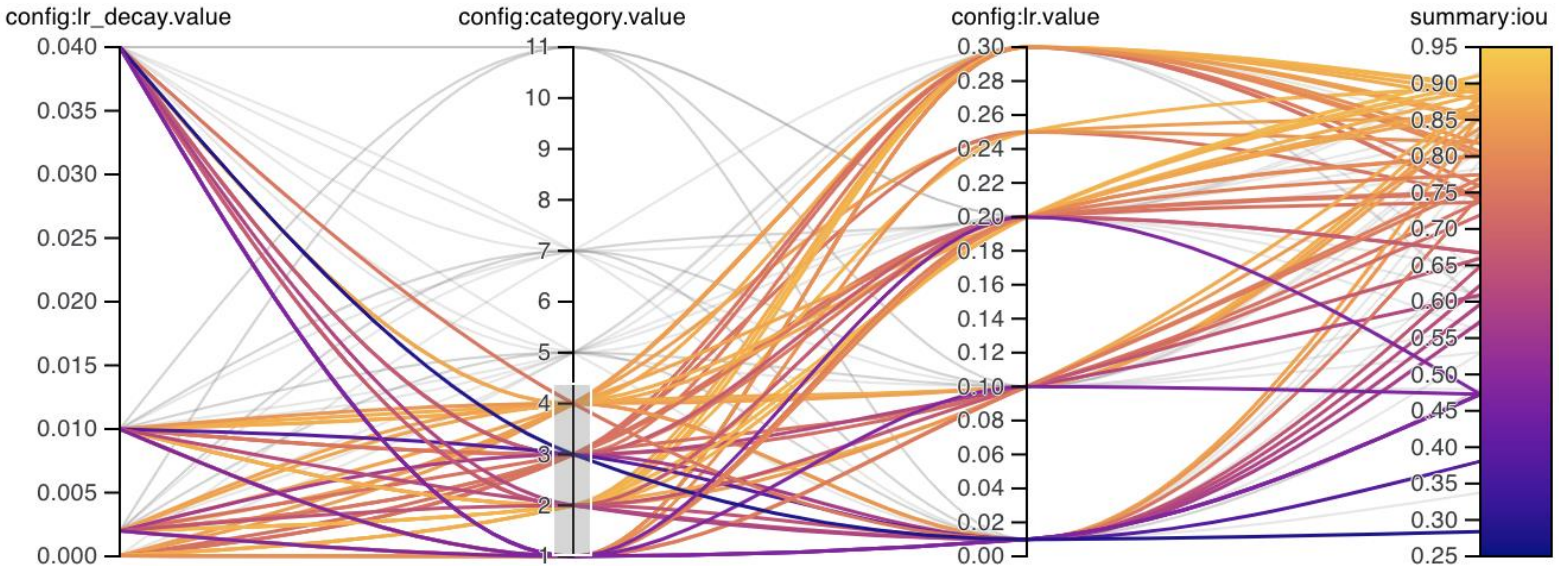
eg. Q: Which model is the best?
↳ (SVM/CNN/RandomForest)

possible Q: compare what has influence on System
7 different stop word lists

Overview


Ab hier automatisch
Train/Test Split
Text Data
Preprocessing
Representation
Model
Cross Validation
-SVM
-LLR
-LNN
↳ label zusammenfassen
↳ Tokenization
↳ Cleaning
↳ Embedding
↳ Word2Vec
↳ BERT
↳ Bag of Words
↳ TF-IDF
Evaluation
↳ F1-Score
↳ Accuracy
Hyper Parameter Tuning
Nächste Woche: 2 Zeilen Research Q formulieren

Hyperparameter Tuning – Parallel Coordinate Plot



https://scikit-learn.org/stable/modules/grid_search.html

Ablation Study

	Baseline	Character n-gram	Counting features	Glove	Word n-gram	Negation	Bag Of Words	Lexica	Last Token Type	Non Cont n-gram	POS n-gram	Swiss-Chocolate Lexicons	Webis
Test 2016	60.30	59.93	60.53	58.01	60.34	60.60	60.00	58.21	60.39	59.99	60.26	60.29	-
Test 2015	63.68	63.21	62.56	61.68	63.34	62.24	63.81	59.46	63.15	63.11	63.18	63.32	64.84
Test 2014	66.57	67.50	67.02	66.00	66.74	68.72	68.13	66.41	66.12	65.75	66.36	66.38	70.86
Test 2013	68.84	68.40	68.15	67.79	68.86	68.72	68.31	66.30	68.75	69.26	68.80	68.43	68.49
Test LiveJournal2014	71.61	70.46	71.46	70.79	70.89	71.35	70.89	65.62	70.68	71.62	71.12	70.67	71.64
Test Tw2014Sarcasm	50.41	50.64	52.98	48.96	53.69	50.00	50.45	54.73	52.59	53.74	51.89	52.62	49.33

Table 4.1: Overall results of the feature inspection. Underlined: the lowest score in the row.

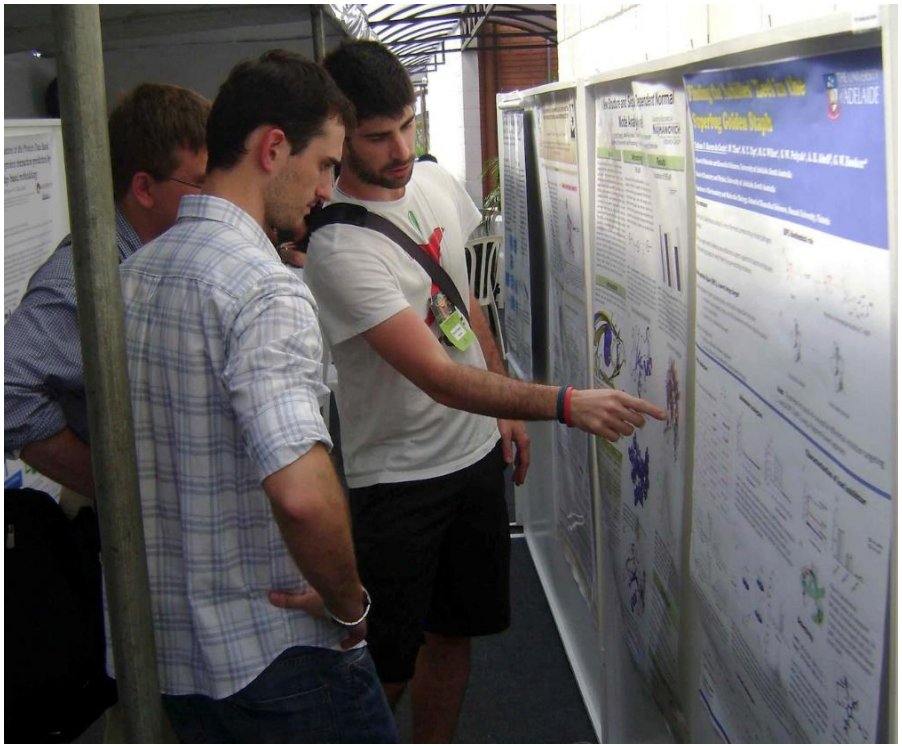
https://www.research-collection.ethz.ch/handle/20.500.11850/155798

Am Schluss Poster Session

Scientific Posters

Poster Session at Scientific Conferences

https://image-factory.media.messe-muenchen.de/image/396372/s1920/



https://fourwaves.com/media/m1qetlzm/poster-vince.jpg?quality=100&rnd=132932157070000000

PIGS IN SPACE: EFFECT OF ZERO GRAVITY AND AD LIBITUM FEEDING ON WEIGHT GAIN IN CAVIA PORCELLUS

Colin B. Purrington
6673 College Avenue, Swarthmore, PA 19081 USA

ABSTRACT:

One ignored benefit of space travel is a potential elimination of obesity, a chronic problem for a growing majority in many parts of the world. In theory, when an individual is in a condition of zero gravity, weight is eliminated. Indeed, in space one could conceivably follow ad libitum feeding and never even gain an gram, and the only side effect would be the need to upgrade one's stretchy pants("exercise pants"). But because many diet schemes start as very good theories only to be found to be rather harmful, we tested our predictions with a long-term experiment in a colony of Guinea pigs (Cavia porcellus) maintained on the International Space Station. Individuals were housed separately and given unlimited amounts of high-calorie food pellets. Fresh fruits and vegetables were not available in space so were not offered. Every 30 days, each Guinea pig was weighed. After 5 years, we found that individuals, on average, weighed nothing. In addition to weighing nothing, no weight appeared to be gained over the duration of the protocol. If space continues to be gravity-free, and we believe that assumption is sound, we believe that sending the overweight — and those at risk for overweight — to space would be a lasting cure.

INTRODUCTION:

The current obesity epidemic started in the early 1960s with the invention and proliferation of elastane and related stretchy fibers, which released wearers from the rigid constraints of clothes and permitted monthly weight gain without the need to buy new outfits. Indeed, exercise today for hundreds of million people involve only the act of wearing stretchy pants in public, presumably because the constrictive pressure forces fat molecules to adopt a more compact tertiary structure (Xavier 1965).

Luckily, at the same time that fabrics became stretchy, the race to the moon between the United States and Russia yielded a useful fact: gravity in outer space is minimal to nonexistent. When gravity is zero, objects cease to have weight. Indeed, early astronauts and cosmonauts had to secure themselves to their ships with seat belts and sticky boots. The potential application to weight loss was noted immediately, but at the time travel to space was prohibitively expensive and thus the issue was not seriously pursued. Now, however, multiple companies are developing cheap extra-orbital travel options for normal consumers, and potential travelers are also creating news ways to pay for products and services that they cannot actually afford. Together, these factors open the possibility that moving to space could cure overweight syndrome quickly and permanently for a large number of humans.

We studied this potential by following weight gain in Guinea pigs, known on Earth as fond of ad libitum feeding. Guinea pigs were long envisioned to be the "Guinea pigs" of space research, too, so they seemed like the obvious choice. Studies on humans are of course desirable, but we feel this current study will be critical in acquiring the attention of granting agencies.

MATERIALS AND METHODS:

One hundred male and one hundred female Guinea pigs (Cavia porcellus) were transported to the International Space Laboratory in 2010. Each pig was housed separately and deprived of exercise wheels and fresh fruits and vegetables for 48 months. Each month, pigs were individually weighed by duct-taping them to an electronic balance sensitive to 0.0001 grams. Back on Earth, an identical cohort was similarly maintained and weighed. Data was analyzed by statistics.

RESULTS:

Mean weight of pigs in space was 0.0000 +/- 0.0002 g. Some individuals weighed less than zero, some more, but these variations were due to reaction to the duct tape, we believe, which caused them to be alarmed push briefly against the force plate in the balance. Individuals on the Earth, the control cohort, gained about 240 g/month ($p = 0.0002$). Males and females gained a similar amount of weight on Earth (no main of effect of sex), and size at any point during the study was related to starting size (which was used as a covariate in the ANCOVA). Both Earth and space pigs developed substantial dewlaps (double chins) and were lethargic at the conclusion of the study.

CONCLUSIONS:

Our view that weight and weight gain would be zero in space was confirmed. Although we have not replicated this experiment on larger animals or primates, we are confident that our result would be mirrored in other model organisms. We are currently in the process of obtaining necessary human trial permissions, and should have our planned experiment initiated within 80 years, pending expedited review by local and Federal IRBs.

ACKNOWLEDGEMENTS:

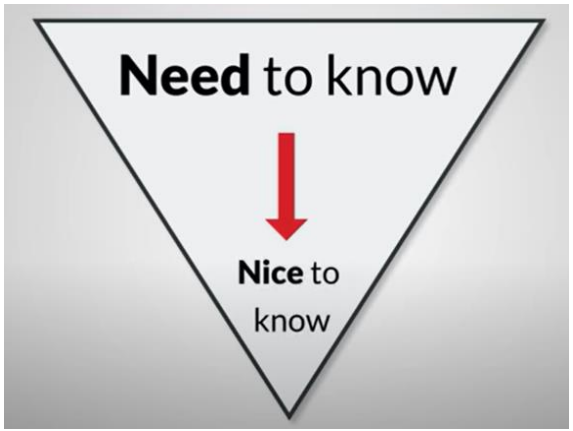
I am grateful for generous support from the National Research Foundation, Black Hole Diet Plans, and the High Fructose Sugar Association. Transport flights were funded by SPACE-EXES, the consortium of wives divorced from insanely wealthy space-flight startups. I am also grateful for comments on early drafts by Mañana Athletic Club, Corpus Christi, USA. Finally, sincere thanks to the Cuy Foundation for generously donating animal care after the conclusion of the study.

LITERATURE CITED:

NASA. 1982. Project STS-XX: Guinea Pigs. Leaked internal memo.
Sekulić, S.R., D. D. Lukač, and N. M. Naumović. 2005. The Fetus Cannot Exercise Like An Astronaut: Gravity Loading Is Necessary For The Physiological Development During Second Half Of Pregnancy. Medical Hypotheses. 64:221-228
Xavier, M. 1965. Elastane Purchases Accelerate Weight Gain In Case-control Study. Journal of Obesity. 2:23-40.

Key Recommendations

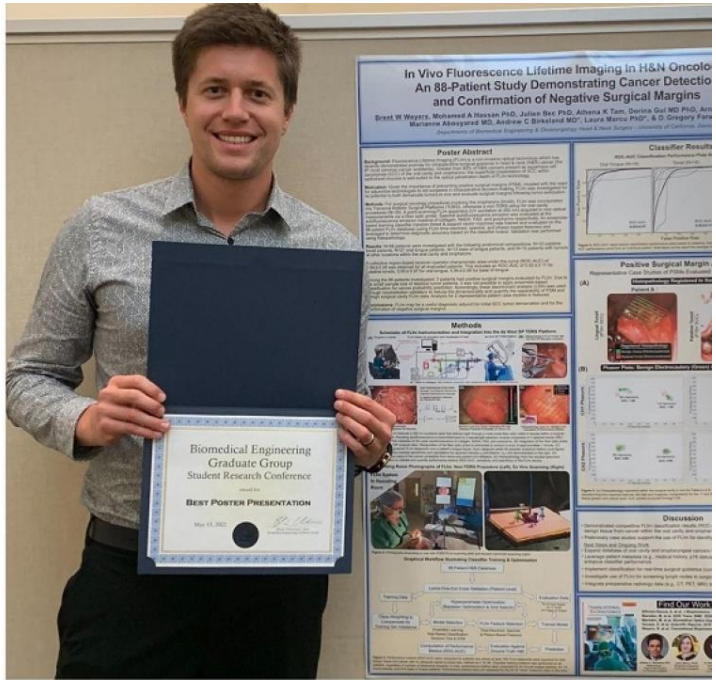
- Descriptive title
- Large font
- As few text as possible
- Landscape mode
- Align boxes
- Few colors
- Clear glow of reading
- Use sans-serif fonts for headlines, serif fonts for text



Poster Presentations

→ siehe Moodle für Beispiel wie Poster sein sollte

- Practice at home
- Know a good opening sentence (e.g. “May i guide you through my research in 2 minutes”)
- Ask audience about their expertise and adapt content
- Present core points in 3 minutes
- Ask for questions
- Finish and allow audience to move on
- Be excited about your work-



https://marculab.bme.ucdavis.edu/sites/g/files/dgvnsk11236/files/styles/sf_landscape_16x9/public/media/images/Bild1.jpg?h=5a210d57&itok=IKKprgE6

Presentation

- on small InnoSwiss project
- needed to be in php bc no one of company can use Python
- chatgpt as solution, bc works for other languages than Python
- Solution: prompts on prompts on prompts with a good output result

Result: better to instruct LLM than telling it what not to do

Q: What type of chatgpt did you use? Enterprise?