L101: Machine Learning for Language Processing

Lecture 8

Guy Emerson

Today’s Lecture

- Problem interpreting results:
  statistical significance

- Problem with datasets:
  social bias

Current State of NLP

- Emphasis on empirical results
- Statistical significance rarely discussed
- Large number of architectures, hyperparameters
- Datasets re-used many times

The large number of architectures and hyperparameters means that testing significance is important (or should be!) – if we test a large number of models, some are bound to perform better than others, just by random variation.

The re-use of datasets (in particular, “standard” datasets like the Penn Treebank) means that the field as a whole may be overfitting, even if each individual paper is not.

Dror et al. (2018) survey

	ACL 2017		TACL 2017	
Total papers	196		37	
Experimental papers	180		33	
– reporting significance	63	(35%)	18	(55%)
– correctly	36	(20%)	15	(45%)

Dror et al. (2018) survey ACL and TACL papers from 2017, and give recommendations for significance testing. `http://aclweb.org/anthology/P18-1128`

Of the papers that report significance incorrectly, some use an inappropriate test (6 ACL papers), and some do not state what test they used (21 ACL papers and 3 TACL papers).

The vast majority of papers are experimental, but significance testing is not the norm!

p-Values

- Probability the result would be at least this extreme, under the null hypothesis

NOT:

- Probability the null hypothesis is true

Most researchers have heard of p-values, but they are often misunderstood!

Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**
- **Significance level**: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)
- **Power**: probability of correctly rejecting null hypothesis (assuming alternative hypothesis)

The null hypothesis formalises the idea that the method doesn't work. The test statistic summarises the results in single number. (How the null hypothesis and test statistic are defined will depend on the task.) If the observed test statistic is too extreme (beyond some threshold), we reject the null hypothesis.

A p-value is a way to re-express the test statistic in terms of a probability. Rather than using the observed test statistic itself, we can calculate the probability that the statistic would be at least as extreme as observed.

In research papers, the term "significant" should be reserved for statistical significance. Some authors use the term loosely, but this is bad practice.

Parametric Tests

- Test statistic follows known distribution (with known parameters)
- Paired Student's t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
  - Test statistic: $t = \frac{\sqrt{n}}{s_D}\bar{x}_D$
  - "Student's t-distribution with $n-1$ degrees of freedom"

The paired Student's t-test is an example of a parametric test. It is appropriate when scores are approximately normally distributed. It is useful when comparing the results of two systems on the same data.

(By the central limit theorem, we can get approximately normally distributed scores by averaging many observations.)

$\bar{x}_D$ is the average difference between the scores of the two systems.
s_D is the standard deviation of the differences between scores.
n is the number of datapoints.

We have to divide by the observed standard deviation, because we don't know what the standard deviation should be. Thed resulting distribution is called Student's t-distribution, and it looks a bit like the normal distribution. The details aren't important here – this is a standard test, available in any reasonable statistics package.

Nonparametric Tests

- No assumptions about distribution
- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: n
  - Binomial distribution

The sign test is an example of a nonparametric test. It is useful when comparing two systems, when we don’t know the distribution of scores – here, we simply look at which system is better.

n is the number of times system A is better than system B.

(In the case of ties, we can evenly split the ties between the two systems, or we can discard them. Discarding them gives a more powerful test – see *power* on slide 5. An alternative is the trinomial test, which includes the ties as a third outcome.)

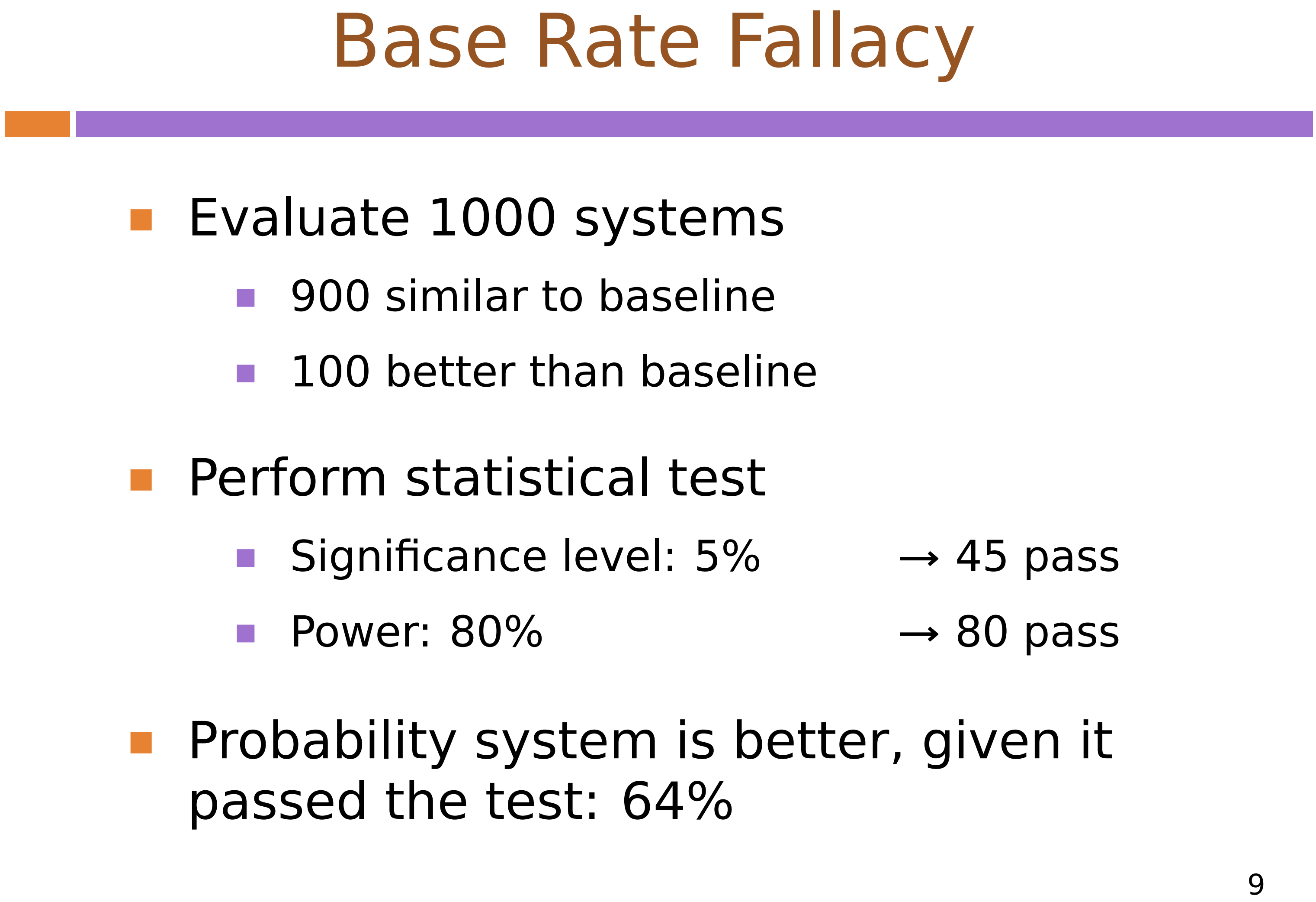
Multiple Tests

- If we test many systems, we expect some will pass
- Bonferroni correction:
  - Replace nominal significance level
  - $\alpha \longmapsto \dfrac{\alpha}{m}$

α is the desired significance level, for all tests combined.
m is the number of systems being tested.
$\frac{\alpha}{m}$ is the significance level that should be used for each individual test.

Further reading:

`https://xkcd.com/882/`



Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline
- Perform statistical test
  - Significance level: 5% → 45 pass
  - Power: 80% → 80 pass
- Probability system is better, given it passed the test: 64%

The base rate fallacy shows why the misunderstanding about p-values is so dangerous.

Here, the probability that the system is better than the baseline, given that it passed the test, is only 64%. This is much lower than 95%!.

The reason for this is the *base rate*, the proportion of tested systems that are actually better.

Base Rate Fallacy

- Evaluate 1000 systems
  - 960 similar to baseline
  - 40 better than baseline
- Perform statistical test
  - Significance level: 5% → 48 pass
  - Power: 80% → 32 pass
- Probability system is better, given it passed the test: 40%

If we reduce the base rate to 4%, the the probability that the system is better than the baseline, given that it passed the test, drops to only 40%.

Base Rate Fallacy

- Evaluate 1000 systems
  - 1000 similar to baseline
  - 0 better than baseline
- Perform statistical test
  - Significance level: 5% → 50 pass
  - Power: 80% → 0 pass
- Probability system is better, given it passed the test: 0%

In the extreme case, a base rate of 0 means that all passes are just due to random variation.

This is not just a toy problem, but a common problem in scientific research. For example, see Ioannidis (2005) “Why Most Published Research Findings are False” `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124`

It’s a problem in machine learning, particularly when researchers test many variants of a system – unless we have good reason to believe that some variants should perform better, the base rate could be very low.

Effect Size

- A significant difference may not be a large difference
- e.g. a coin toss
  - Coins not perfectly symmetric
  - Probability of heads not exactly 50%
  - Difference so small we don’t care

Publication Bias

- Hard to publish negative results...
- Authors may hide failed experiments
- MPhil project and L101 mini-project: Don't hide! Negative results are okay!

Publication bias means that we get a skewed view of results. Remember that when we make multiple tests, we need to correct for this, e.g. using the Bonferroni correction. However, if negative results are not published, we don't get to see how many experiments were run.

This becomes more serious when publication bias leads to authors changing how they try to present their work.

For your MPhil, you don't have to worry about a reviewer misguidedly demanding positive results! Try to run your experiments carefully, and report whatever you find.

Summary of Significance Testing

- Significance testing is important but underused in NLP!
- Choice of test:
  - Parametric (e.g. paired Student’s t-test)
  - Nonparametric (e.g. sign test)
  - Multiple tests (e.g. Bonferroni correction)
- Be careful:
  - Base rate fallacy
  - Effect size
  - Publication bias

Besides significance testing, there are many other methodological issues not discussed here. For example, variance in results (so use multiple runs!), correlation between datapoints, error analysis, human evaluation.

For advice on significance testing (in Cambridge), the Statistical Laboratory runs a free statistics clinic:
`http://www.statslab.cam.ac.uk/clinic/`

Further reading:
Reinhart (2015)
`https://www.statisticsdonewrong.com/`
Søgaard et al. (2014)
`http://aclweb.org/anthology/W14-1601`
Koehn (2004)
`http://aclweb.org/anthology/W04-3250`
Berg-Kirkpatrick et al. (2012)
`http://aclweb.org/anthology/D12-1091`
Faruqui et al. (2016)
`http://aclweb.org/anthology/W16-2506`

Back to the Beginning...

- Task
- Data

What if this goes wrong?

- Model
- Training

Most NLP papers

- Real-world application?

Recall how we split up machine learning in the first lecture.

If the task is poorly defined, or there is a problem with the data, any machine learning model is going to struggle.

And if something goes wrong, what happens if we use the trained system in a real-world application?

Caruana et al. (2015)

- Task: Predict death from pneumonia
- Pattern in data: asthma reduces risk
- Real reason: asthma patients sent to Intensive Care Unit, reducing risk
- Shallow models (e.g. logistic regression) → can identify and fix such problems

An example from healthcare, to demonstrate the problem – this example is serious but uncontroversial. Patients with a high risk of death would be treated in the hospital, while patients with a low risk would be treated as outpatients. If a high-risk patient is mistakenly sent home, and then they die, this is a serious mistake.

Caruana et al. (2015) show how a real pattern in the data is having asthma correlates with lower risk – despite asthma and pneumonia both being lung conditions. This is because the asthma patients in the dataset were in fact given intensive care, and improved as a result of that care. Here, there is a bias in the dataset that we don't want in the trained model. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.704.9327&rep=rep1&type=pdf`

Bias

- Bias (statistics):
  expected value differs from true value
- Bias (law):
  unfair or undesirable prejudice

There is nothing morally wrong with statistical bias.

In this lecture, I’m discussing the day-to-day or legal meaning of “bias”.

Bias

“Bias is a social issue first, and a technical issue second.”

(Crawford, 2017)

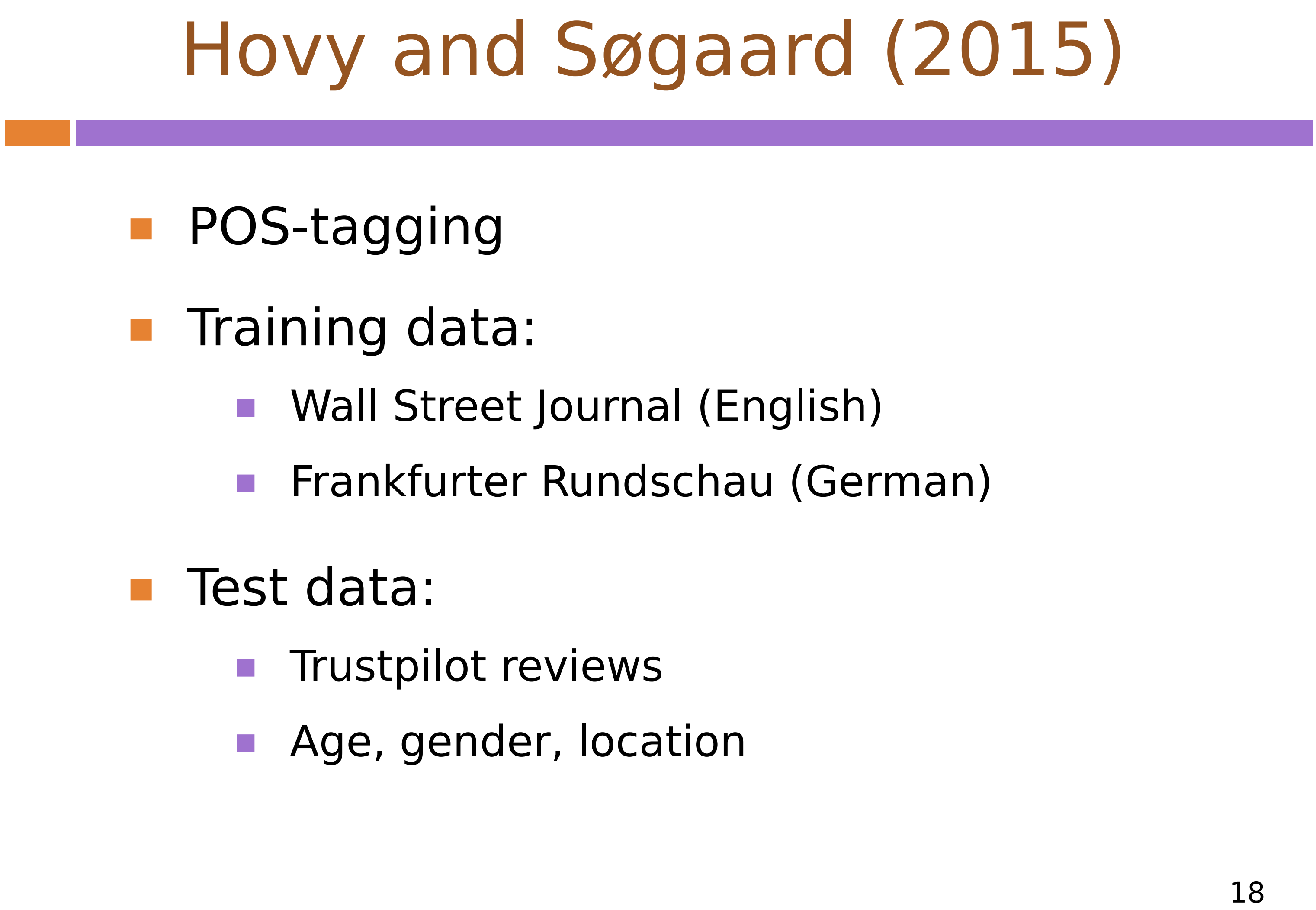
Kate Crawford (2017) “The Trouble With Bias”, NIPS Keynote Lecture
`https://www.youtube.com/watch?v=ggzWIipKraM`

Many machine learning researchers prefer to work on technical issues, but the social issues are still there. Social issues are important for any real-world application.

Demographic Bias

- Region
- Social Class
- Gender
- Age
- Ethnicity

NLP doesn’t have life-and-death examples like in healthcare, but there are still important biases to consider. Do NLP tools work equally well for all demographic groups?



Hovy and Søgaard (2015)

- POS-tagging
- Training data:
  - Wall Street Journal (English)
  - Frankfurter Rundschau (German)
- Test data:
  - Trustpilot reviews
  - Age, gender, location

http://aclweb.org/anthology/P15-2079

Part-of-speech tagging is often considered a solved problem, or at least nearly-solved. However, is it solved for all demographic groups?

The training datasets are standard datasets.

The test data already differs in genre (reviews rather than newspapers) and in domain (different subject matters). As we will see, the performance is overall lower than we might expect from the state of the art, when evaluating on the same genre and domain as the training data.

Importantly, because each review is associated with a user, we have access to demographic metadata.

H&S (2015) – German Results

Group	TreeT	CRF++
Under 35	.874	.859
Over 45	**.894**	.870
Men	.885	.861
Women	.882	**.868**
Highest-prob region	.885	.865
Lowest-prob region	.889	.874

The bold results are statistically significant. (As we've seen, given the current state of NLP, this is almost a luxury.) Two passes out of six is more than we'd expect.

For age, the effect size is between 1 and 2% – this is the kind of difference that many papers will point to when introducing a new model. Here, we can see this kind of improvement is wiped out just by changing the age group.

For gender, the effect is much smaller (and in the opposite direction for each tagger). Note how there is a significant difference for CRF++, even though the difference in performance is smaller than for age. Remember – significance and effect size are not the same!

It is good that results on regions were published, even though this is a negative result. Here, "prob" refers to the probability assigned by the model. A probabilistic model can estimate the probability of any input sequence – it seems plausible that if the model judges a sequence to be more likely, it will be also be more accurate. However, that was not the case here.

H&S (2015) – English Results

Group	TreeT	CRF++
Under 35	.879	.882
Over 45	**.883**	**.884**
Men	.882	.886
Women	.880	.881
Highest-prob region	.883	.886
Lowest-prob region	.882	.885

All differences are small, but for age, the difference is significant.

Jørgensen et al. (2015)

POS-tagging on Twitter data

Group	Stanf.	Gate	Ark
AAVE	.614	.791	.775
non-AAVE	**.745**	**.833**	.779

`http://aclweb.org/anthology/W15-4302`

Jørgensen et al. look at African American Vernacular English (AAVE).

For two of the taggers, the effect size is substantial.

For AAVE, PoS-tagging is far from a solved problem!

The Gate and Ark taggers have been adapted for Twitter, while the Stanford tagger is not (but is often treated as a standard tool).

Caliskan et al. (2017)

- Corpora reflect social biases:
  - Uncontroversial (e.g. pleasant/unpleasant association with flowers, insects, etc.)
  - Prejudiced (e.g. pleasant/unpleasant association with gender, ethnicity, etc.)
  - Status quo (e.g. association between gender and career)
- Distributional semantic vectors reflect social biases

`https://purehost.bath.ac.uk/ws/portalfiles/portal/168480066/CaliskanEtAl_authors_full.pdf`

Caliskan et al. (2017) look at distributional vector space models, and measure correlations with real-world measurements. They use the implicit association test (a well-known psychological test) to measure associations that human participants have, and show that these correlate with distributional similarity. For careers, they use figures from the US Bureau of Labor, and show that these also correlate with distributional similarity.

Given that these associations exist in our culture, it is perhaps unsurprising that these associations are discovered through distributional semantics. However, if such a system is then used to make real-world decisions, the danger is that the associations are not just observed, but *reproduced*.

(The discussion of gender is relevant for machine learning, which currently has a strong skew towards men. The situation in NLP is more balanced, but still not perfect.)

Decision Making

- The Guardian (2017): “Computer says no: Irish vet fails oral English test needed to stay in Australia”
- Bias in training data vs. bias in decisions

Guardian article:
`https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia`

Follow-up Guardian article:
`https://www.theguardian.com/australia-news/2017/aug/10/outsmarting-the-computer-the-secret-to-passing-australias-english-proficiency-test`

This is a newspaper article, not a research article, so there’s no comparison between English speakers from different countries. However, it illustrates the point that NLP tools are being used in practice, sometimes without carefully considering how they might go wrong.

Regardless of whether this particular system has a problem with Irish accents, this is a plausible problem. In a real-world application, we need to make sure that a bias in the training data (such as not having any Irish people) doesn’t result in biased decisions (such as rejecting visas for Irish people).

Summary of Bias and Ethics

- Social bias (not statistical bias)
  - Training data
  - Model predictions
- POS-tagging & demographic groups
- Distributional semantics & associations

Further reading:

Hovy and Spruit (2016) `http://aclweb.org/anthology/P/P16/P16-2096.pdf`

The ACL wiki has a page listing online resources from courses on ethics in NLP. `https://aclweb.org/aclwiki/Ethics_in_NLP`

Finally, Widening NLP (WiNLP) is a group within the ACL community, which aims to support underrepresented groups. They have organised an annual workshop since 2017. `http://www.winlp.org/`

Course Summary

- Naive Bayes, Topic Classification
- HMM, POS-Tagging
- Logistic Regression, MEMM, NER
- Decision Boundaries, SVM, Kernels
- K-Means, LDA, WSI, Topic Discovery
- Distributional Semantics
- CNN, RNN, Hyperparameter Tuning
- Statistical Significance, Social Bias

Still To Come

- Last 3 sessions – reading seminars
- Mini-project